VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Využití metod časoprostorové analýzy v praxi

Exploratory Spatio–Temporal Analysis in pracitse

2014 Lenka Vozárová

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zadání bakalářské práce

Student: **Lenka Vozárová**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 1103R031 Výpočetní matematika

Téma: Využití metod časoprostorové analýzy v praxi
Exploratory Spatio-Temporal Analysis in practise

Zásady pro vypracování:

S rostoucím využíváním informačních technologií v běžném životě obyvatel rostou i objemy jimi průběžně získávaných časoprostorových dat. Cílem práce je seznámit se s principy a metodami časoprostorové analýzy, naučit se volit vhodné postupy, prakticky je aplikovat a kriticky hodnotit získané výsledky.

Postup práce:
1. Seznámení se s metodami časoprostorové analýzy dat.
2. Časoprostorová analýza požárů ve vybraném regionu.

Seznam doporučené odborné literatury:

1. HAMILTON, J. (1994), Time Series Analysis, Princeton: Princeton Univ. Press, ISBN 0-691-04289-6.
2. PECÁKOVÁ, I. (2011), Statistika v terénních průzkumech, Praha: Professional Publishing, ISBN: 978-80-7431.
3. ŘEZÁNKOVÁ, H. (2005), Analýza kategoriálních dat, Praha: Oeconomica, ISBN: 80-245-0926-1
4. LITSCHMANNOVÁ, M. (2012), Úvod do statistiky, Učební texty v elektronické podobě, dostupné z Word Wide Web: http://mi21.vsb.cz/modul/uvod-do-statistiky

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Ing. Martina Litschmannová, Ph.D.**

Datum zadání: 01.09.2013

Datum odevzdání: 07.05.2014

VYSOKÁ ŠKOLA BÁŇSKÁ - TECHNICKÁ UNIVERZITA OSTRAVA * Fakulta elektrotechniky a informatiky *

doc. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prehlasujem, že som túto bakalársku prácu vypracovala samostatne. Uviedla som všetky literárne pramene a publikácie, z ktorých som čerpala.

V Ostrave 7. mája 2014 .....

Na tomto mieste by som chcela poďakovať všetkým, ktorí ma počas písania práce podporovali. Veľká vďaka patrí vedúcej bakalárskej práce Ing. Martine Litschmannovej, Ph.D. za venovaný čas, trpezlivosť a odborné rady. Ďalej ďakujem riaditeľovi OR HaZZ v Spišskej Novej Vsi Mgr. Martinovi Vozárovi za poskytnuté dáta a cenné informácie.

Abstrakt

Táto bakalárska práca pomocou metód časopriestorovej analýzy dát skúma vývoj počtu požiarov v okresoch Spišská Nová Ves a Gelnica. Prvá polovica práce sa zaoberá teoretickým popisom štatistických metód, ktoré sa venujú najmä analýze kategoriálnych dát. Ich praktické využitie je opísané v druhej časti práce.

Kľúčové slová: ANOVA, kontingenčné tabuľky, časové rady

Abstract

This bachelor thesis by exploratory spatio-temporal analysis examine development of number of fires in district Spišská Nová Ves and Gelnica. First half of thesis deals with teoretical description of statistical methods, which address mainly to analysis of categorical data. Their practical use is described in second part of thesis.

Keywords: ANOVA, contingency tables, time series

Zoznam použitých skratiek a symbolov

ANOVA	–	Analýza rozptylu (z anglického *Analysis of variance*)
B	–	Testová štatistika Bartlettovho testu
C_p	–	Pearsonov koeficient kontingencie
F	–	Testová štatistika F-pomer
$F(x)$	–	Distribučná funkcia
H_A	–	Alternatívna hypotéza
H_0	–	Nulová hypotéza
$N(\mu, \sigma^2)$	–	Normálne rozdelenie so strednou hodnotou μ a rozptylom σ^2
OR HaZZ	–	Okresné riaditeľstvo Hasičského a záchranného zboru
P	–	Pearsonova štatistika chí-kvadrát
S_B^2	–	Medzitriedny výberový rozptyl
S_W^2	–	Vnútorný výberový rozptyl
SS_B	–	Mezitriedna variabilita
SS_T	–	Totálna variabilita
SS_W	–	Vnútorná variabilita
V	–	Cramerov koeficient kontingencie
W	–	Testová štatistika Shapirovho–Wilkovho testu
X	–	Náhodná veličina
Y	–	Testová štatistika Yatesovej korekcie χ^2 testu o nezávislosti
k	–	Počet jednotlivých výberov
$\bar{k}$	–	Priemerný koeficient rastu
n	–	Celkový rozsah výberu
x_{OBS}	–	Pozorovaná hodnota testovej štatistiky
α	–	Hladina významnosti testu
β	–	Pravdepodobnosť chyby II. druhu.
$\bar{\Delta}$	–	Priemerný absolútny prírastok
$\bar{\delta}$	–	Priemerný relatívny prírastok
μ	–	Stredná hodnota
σ^2	–	Rozptyl

Obsah

Zoznam tabuliek

Zoznam obrázkov

1 Úvod

Bakalárska práca sa zaoberá využívaním metód časopriestorovej analýzy v praxi. Tieto metódy sa začali využívať už na prelome 19. a 20. storočia. Vďaka rozmachu informačných technológií sa tešia čoraz väčšej popularite.

Túto tému som si vybrala z dôvodu, že obaja rodičia pracujú v Hasičskom a záchrannom zbore, teda k požiarom a všeobecne hasičom mám blízko už od detstva. Zo štatistického hľadiska ma vždy zaujímalo, kde a kedy je najvyšší počet požiarov, aké sú ich najčastejšie príčiny a od čoho závisia. Práve preto bola voľba tejto bakalárskej práce pre mňa ideálna.

Dáta k analýze som získala od OR HaZZ v Spišskej Novej Vsi. Ten sa delí na dva okresy (Spišská Nová Ves a Gelnica), v ktorých sa nachádzajú tri hasičské stanice (Spišská Nová Ves, Gelnica a Krompachy). Pod OR HaZZ Spišská Nová Ves patrí 56 obcí, s počtom obyvateľov takmer 129 000 (k máju 2011). Skúmané dáta pochádzajú z obdobia medzi rokmi 2003 - 2012.

Práca je rozdelená na teoretickú a praktickú časť. V kapitolách 2 - 4 sú vysvetlené základné pojmy analýzy rozptylu, kontingenčných tabuliek a časových radov, ktoré budú využívané v praktickej časti. Cieľom praktickej časti v kapitolách 5 - 7 je skúmanie časových a priestorových závislostí medzi požiarmi a ich vývoj za určité obdobie. Jedná sa hlavne o priebeh požiarov počas týždňa, roka a závislosť počtu požiarov podľa príčiny na zásahovom obvode.

2 Testovanie hypotéz

Na úvod si v skratke pripomeňme o čo sa jedná pri testovaní hypotéz. Cieľom je overiť, že dáta nepopierajú predpoklad, ktorý sme si stanovili pred začiatkom testovania. Hypotézy rozdeľujeme na **parametrické** (hypotézy o parametroch rozdelenia) a **neparametrické** (hypotéza o iných vlastnostiach rozdelenia). U oboch rozhodujeme medzi nulovou a alternatívnou hypotézou:

Nulová hypotéza H_0 : tvrdenie, predstavujúce akýsi rovnovážny stav.

Alternatívna hypotéza H_A : tvrdenie, ktoré popiera nulovú hypotézu.

Pri rozhodovaní môžu nastať 4 situácie, ktoré popisuje tabuľka 2.1.

	Nezamietame H_0	**Zamietame** H_0
Platí H_0	Správne rozhodnutie pravdepodobnosť: $1-\alpha$	**Chyba I. druhu** pravdepodobnosť: α
Platí H_A	**Chyba II. druhu** pravdepodobnosť: β	Správne rozhodnutie pravdepodobnosť: $1-\beta$

Tabuľka 2.1: Výsledky testovania hypotéz

Pravdepodobnosť α nazývame hladina významnosti testu a pravdepodobnosť $1-\beta$ zvykne byť označovaná ako tzv. sila testu [2]. V praxi väčšinou volíme hladinu významnosti α ako 0,05 (resp. 0,01).

K testovaniu hypotéz môžeme pristupovať dvoma spôsobmi – **klasickým testom** alebo **čistým testom významnosti**. Nevýhodou klasického testu je, že na začiatku testovania musíme určiť kritický obor, ktorý sa mení vplyvom hladiny významnosti α . Naopak, pri čistom teste významnosti rozhodujeme pomocou **p-hodnoty**, čo je najnižšia hladina významnosti, na ktorej môžeme H_0 zamietnuť. Ďalej sa budeme zaoberať len čistým testom významnosti, ktorý sa skladá z následujúcich krokov:

1. Formulácia nulovej a alternatívnej hypotézy
2. Voľba testovej štatistiky
3. Overenie predpokladov testu
4. Výpočet pozorovanej hodnoty x_{OBS} testovej štatistiky
5. Výpočet p-hodnoty
6. Rozhodnutie o výsledku testu

Posledné dva kroky pre jednoduchosť môžeme zapísať do tabuliek 2.2 a 2.3.

Tvar H_A	$\boldsymbol{p - hodnota}$
$\theta < \theta_0$	$F_0(x_{OBS})$
$\theta > \theta_0$	$1 - F_0(x_{OBS})$
$\theta \neq \theta_0$	$2\ min\{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$

Tabuľka 2.2: Výpočet p-hodnoty pri testovaní parametrických hypotéz
[1]

V tabuľke 2.2 je θ testovaný parameter a θ_0 je očakávaná hodnota.

$\boldsymbol{p - hodnota}$	**Rozhodnutie**
$p - hodnota < 0,01$	H_0 zamietame v prospech H_A
$0,01 < p - hodnota < 0,05$	Doporučujeme opakovať test s väčším rozsahom výberu
$p - hodnota > 0,05$	H_0 nezamietame

Tabuľka 2.3: Rozhodnutie pomocou p-hodnoty
[1]

2.1 ANOVA

Analýza rozptylu (ANOVA) je rozšírením dvojvýberových testov strednej hodnoty. Umožňuje nám zrovnávať niekoľko stredných hodnôt nezávislých náhodných výberov. Táto metóda v porovnaní s opakovanými dvojvýberovými t–testmi zachováva rozumnú silu testu a hladinu významnosti α.

ANOVA vo svojej parametrickej podobe musí splňovať určité predpoklady. Aj keď bola pôvodne navrhnutá pre rovnaký rozsah jednotlivých výberov nie je to predpokladom testu. Napriek tomu platí, že čím viac je toto pravidlo splnené, tým presnejších výsledkov dosiahneme. Jednotlivé predpoklady budú popísané v časti 2.1.1.

Dôvodom podrobného zaoberania sa analýzou rozptylu je, že v praktickej časti (kapitola 5) budem skúmať počet požiarov v priebehu týždňa.

2.1.1 Predpoklady testu ANOVA a ich overenie

Medzi základné predpoklady analýzy rozptylu patrí

- normalita rozdelení
- homoskedasticita (identické rozptyly)

Ak nie sú splnené tieto predpoklady, môžeme použiť neparametrický *Kruskalov-Wallisov test* [1], nazývaný aj neparametrická ANOVA. Je to viacvýberový test zhody mediánov, ktorý nepredpokladá normalitu rozdelenia.

Keď už poznáme predpoklady testu, ostáva nám ich overenie. Na to slúži množstvo testov. Na overenie normality rozdelenia sa používa napríklad χ^2–test dobrej zhody, modifikovaný Kolmogorovov–Smirnovov test alebo Shapirov–Wilkov test, ktorý využijeme

aj my. Homoskedasticitu overíme napríklad pomocou Bartlettovho testu alebo v prípade nesplnenia normality dát pomocou Leveneovho testu. Testy, ktoré budem využívať v praktickej časti rozoberiem podrobnejšie, ostatné testy sú popísané napr. v [1].

Shapirov–Wilkov test

V tomto prípade testujeme nulovú hypotézu, či výber $x_1, \cdots, x_n$ pochádza z normálneho rozdelenia.
Testová štatistika má tvar

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.1}$$

kde $\boldsymbol{x} = (x_1, \ldots, x_n)$ je vektor usporiadaných náhodných pozorovaní, $\bar{x}$ je ich výberový priemer a a_i sú tabelované váhy pre vybrané rozsahy.

Nulovú hypotézu na hladine významnosti α zamietame ak W je menšie ako tabelovaná kritická hodnota. Podrobnejšie informácie a potrebné tabuľky sa nachádzajú napr. v [3].

Bartlettov test

Bartlettov test je citlivý na porušenie normality. Keďže moje dáta (požiare) skúmané v kapitole 5 pochádzajú z normálneho rozdelenia budem na overenie homoskedasticity používať práve tento test.

Testujeme nulovú hypotézu, že všetky k rozptyly sú si rovné oproti alternatíve, že aspoň 2 z nich sa líšia.
Nech $MS_e = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2$ a $C = 1 - \frac{1}{a(k-1)}\left(\frac{1}{n-k} - \sum_{i=1}^{k}\frac{1}{n_i - 1}\right)$.

Testová štatistika má tvar

$$B = \frac{1}{C}[(n-k)lnMS_e - \sum_{i=1}^{k}(n_i - 1)lns_i^2], \tag{2.2}$$

kde n je celkový rozsah výberu, k je počet jednotlivých výberov, n_i je počet pozorovaní v i-tom výbere a s_i^2 je výberový rozptyl i-tého výberu.

P-hodnotu vypočítame podľa vzťahu $1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribučná funkcia rozdelenia χ^2 s $n - k$ stupňami voľnosti.

2.1.2 F–pomer

Predpokladajme k nezávislých náhodných výberov, ktoré splňujú predpoklady:

$(X_{11}, X_{12}, \ldots, X_{1n_1}) \to N(\mu_1, \sigma^2)$
$(X_{21}, X_{22}, \ldots, X_{2n_2}) \to N(\mu_2, \sigma^2)$
...
$(X_{k1}, X_{k2}, \ldots, X_{kn_k}) \to N(\mu_k, \sigma^2)$,
kde n_i je počet pozorovaní v i-tom výbere, μ_i je stredná hodnota i-tého výberu a σ^2 je rozptyl.

Testujeme hypotézu $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ oproti alternatíve $H_A : \neg H_0$.

Aby sme mohli rozhodnúť, či nulovú hypotézu zamietame alebo nezamietame, musíme nájsť vhodnú testovú štatistiku.

Definujme **vnútornú variabilitu**

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \tag{2.3}$$

medzitriednu variabilitu

$$SS_B = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2, \tag{2.4}$$

vnútorný výberový rozptyl

$$S_W^2 = \frac{SS_W}{n-k}, \tag{2.5}$$

medzitriedny výberový rozptyl

$$S_B^2 = \frac{SS_B}{k-1}. \tag{2.6}$$

Kde platí, že

$n = \sum_{i=1}^{k} n_i$ je celkový počet pozorovaní,

$\bar{X}_i$ je výberový priemer i-tého náhodného výberu,

$\bar{X}$ je celkový výberový priemer.

Súčet vnútornej a medzitriednej variability môžeme definovať ako **totálnu variabilitu**, ktorá má tvar

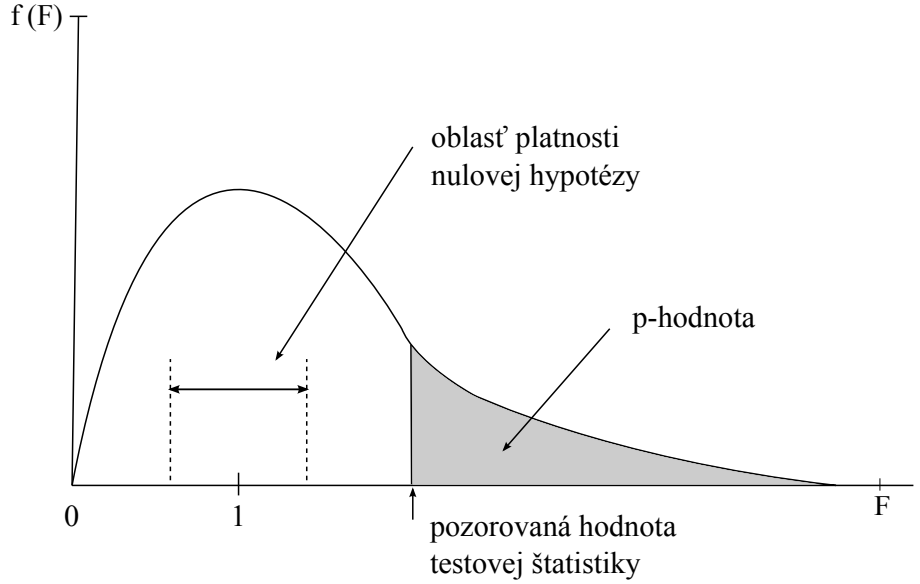
$$SS_T = SS_W + SS_B = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2. \tag{2.7}$$

Teraz môžeme hľadanú testovú štatistiku F nazývanú **F–pomer** definovať ako

$$F = \frac{S_B^2}{S_W^2}. \tag{2.8}$$

P-hodnotu vypočítame podľa vzťahu $1 - F_0(x_{OBS})$, kde $F_0(x)$ distribučná funkcia Fisherovho-Snedecorovho rozdelenia s $k-1$ stupňami voľnosti v čitateli a $n-k$ stupňami voľnosti v menovateli.

Pri rozhodovaní o platnosti hypotézy H_0 využijeme štatistické správanie F-pomeru, čo ilustruje nasledujúci obrázok 2.1.



Obrázok 2.1: Ilustrácia p-hodnoty pre F-pomer

Pre jednoduchosť a prehľadosť sa výsledky výpočtov systematicky zapisujú do tabuľky ANOVA (viď tabuľka 2.4).

Zdroj premenlivosti	Variabilita	Stupne voľnosti	Rozptyl	F-pomer	p-hodnota
Medzitriedny	$SS_B = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$	$k-1$	$S_B^2 = \frac{SS_B}{k-1}$	$F = \frac{S_B^2}{S_W^2}$	$1-F_0(x_{OBS})$
Vnútorný	$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij}-\bar{X}_i)^2$	$n-k$	$S_W^2 = \frac{SS_W}{n-k}$		
Celkový	$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$	$n-1$			

Tabuľka 2.4: Tabuľka ANOVA

2.2 Post hoc analýza

Ak nezamietame H_0, znamená to, že dáta nemajú rozdielne stredné hodnoty a týmto testovanie končí. V prípade zamietnutia H_0 v prospech H_A, je potrebné určiť ktoré výbery sa od seba štatisticky významne líšia. Tento proces nazývame **post hoc analýza.** Jej základom je porovnávanie stredných hodnôt všetkých dvojíc, t.j. testovanie hypotéz

$$H_0 : \mu_i = \mu_j \quad \text{voči} \quad H_A : \mu_i \neq \mu_j,$$

kde pre každé i, j platí $i \neq j$.

Pre tieto tzv. viacnásobné porovnávania existuje viacero metód ako Tukeyho alebo Scheffého metóda popísané napr. v [4]. V našom prípade sa budeme zaoberať konkrétne metódou TukeyHSD.

TukeyHSD

TukeyHSD je modifikáciou klasickej Tukeyho metódy. Využíva sa najmä pri nevyvážených triedeniach. Nulovú hypotézu zamietame ak

$$\mid \bar{x_i} - \bar{x_j} \mid \geq q_a(k, n-k)\sqrt{S_W^2}\sqrt{\frac{1}{2}(\frac{1}{n_i} + \frac{1}{n_j})}, \tag{2.9}$$

kde $q_a(k, n-k)$ je α kvantil tabelovaného studentizovaného rozpätia.

Keď určíme, ktoré výbery sa od seba štatisticky významne líšia mali by sme to prehľadne zobraziť. Medzi základné metódy **prezentovania výsledkov post hoc analýzy** patrí **znamienková schéma** a **homogénne skupiny**.

V znamienkovej schéme sa dvojice, ktoré sa štatisticky významne líšia zobrazujú pomocou určitého symbolu (napr. krížik) do tabuľky o veľkosti $k \times k$, kde k je počet výberov.

U homogénnych skupín (skupiny, u ktorých nebola zamietnutá nulová hypotéza) sa porovnávané výbery vzostupne zoraďujú do tabuľky podľa výberového priemeru. Overovanie zhody medzi prvou a ostatnými skupinami trvá dovtedy, kým nezamietame nulovú hypotézu. Takýmto spôsobom dostaneme prvú homogénnu skupinu, ďalšie získame rovnakým postupom.

3 Kontingenčné tabuľky

3.1 Základné pojmy

Keďže sa v mojej bakalárskej práci budem zaoberať aj závislosťou príčiny požiaru na zásahovom obvode (kapitola 6), za najvhodnejšiu metódu skúmania považujem práve analýzu dát zapísaných v kontingenčnej tabuľke.

Závislosť resp. nezávislosť dvoch kategoriálnych premenných zisťujeme pomocou dvojrozmernej tabuľky početností, ktorú nazývame **kontingenčná tabuľka**. Táto tabuľka vzniká postupným roztriedením prvkov výberu podľa variantov dvoch kategoriálnych znakov. Označme si tieto znaky napr. znak X, ktorý má a variantov a znak Y, ktorý má b variantov. V hlavičke tabuľky sú uvedené názvy variantov znakov X,Y a na okraji tabuľky marginálne početnosti označené $n_{i.}$, $n_{.j}$. Podstatu tabuľky tvoria absolútne početnosti n_{ij}, pre ktoré platí $i \in \langle 1, a\rangle$ a $j \in \langle 1, b\rangle$. Pre lepšiu predstavu si uveďme schému kontingenčnej tabuľky.

X/Y	Y_1	$\cdots$	Y_j	$\cdots$	Y_b	Spolu
X_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1b}	$n_{1.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ib}	$n_{i.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_a	n_{a1}	$\cdots$	n_{aj}	$\cdots$	n_{ab}	$n_{a.}$
Spolu	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.b}$	n

Tabuľka 3.1: Schéma kontingenčnej tabuľky

Okrem vyššie uvedených prvkov sa môžu do kontingenčnej tabuľky zapisovať aj:

- **relatívne početnosti** – podiel príslušnej absolútnej početnosti a celkového rozsahu výberu
- **riadkové relatívne početnosti** – podiel príslušnej absolútnej početnosti a marginálnej početnosti v príslušnom riadku
- **stĺpcové relatívne početnosti** – podiel príslušnej absolútnej početnosti a marginálnej početnosti v príslušnom stĺpci

Podrobnejšie informácie sa nachádzajú napr. v [1].

Na zobrazenie závislosti medzi kategoriálnymi premennými sa používajú aj grafy. Medzi základné patrí **mozaikový graf**, **100% skladaný graf** alebo **zhlukový graf**. Detailnejšie sa k týmto grafom a praktickému využitiu kontingenčných tabuliek vrátime v kapitole 6.

3.2 Testy v kontingenčnej tabuľke

Základným testom pri zisťovaní vzájomnej závislosti dvoch kategoriálnych premenných je **χ^2 test o nezávislosti.** Okrem neho sa využíva aj **vierohodnostný pomer** [5]. Pri nesplnení predpokladov testu, ktoré budú uvedené v 3.2.1 sa používa **Yatesova korekcia χ^2 testu o nezávislosti**.

3.2.1 χ^2 test o nezávislosti

Princíp spočíva v testovaní zhody zistených (n_{ij}) a očakávaných početností (m_{ij}), ktoré vypočítame ako

$$m_{ij} = \frac{n_{i.}n_{.j}}{n}, \tag{3.1}$$

kde $n_{i.}$ sú riadkové marginálne početnosti, $n_{.j}$ sú stĺpcové marginálne početnosti a n je celkový rozsah výberu.
Predpokladom pre využitie testu je splnenie podmienok:

- očakávané početnosti m_{ij} neklesnú pod 2,
- aspoň 80% očakávaných početností m_{ij} je väčších ako 5.

Ak sú splnené predpoklady testu, môžeme testovať nulovú a alternatívnu hypotézu, ktoré majú tvar
H_0 : Znaky X a Y v kontingenčnej tabuľke sú nezávislé.
H_A : Znaky X a Y v kontingenčnej tabuľke sú závislé.

Ako testovú štatistiku použijeme **Pearsonovu štatistiku chí-kvadrát**, ktorú vyjadríme vzťahom

$$P = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{3.2}$$

P-hodnotu vypočítame podľa vzťahu $1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribučná funkcia rozdelenia χ^2 s $(a-1)(b-1)$ stupňami voľnosti.

3.2.2 Yatesova korekcia χ^2 testu o nezávislosti

Ako už bolo spomenuté v úvode tejto kapitoly, táto korekcia sa využíva ak nie sú splnené predpoklady χ^2 testu o nezávislosti (tzn. že máme veľmi nízke očakávané početnosti). V tomto teste znižujeme hodnotu testovej štatistiky, čo má za následok obtiažnejšie zamietnutie nulovej hypotézy. Oproti χ^2 testu o nezávislosti sa líši len v testovej štatistike, ktorá má tvar

$$Y = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij} - m_{ij} - 0,5)^2}{m_{ij}}. \tag{3.3}$$

Výpočet p-hodnoty je rovnaký ako v predošlom prípade.

3.3 Miery kontingencie

Testy opísané v kapitole 3.2 nám po zamietnutí nulovej hypotézy nehovoria nič o intenzite závisloti medzi znakmi X a Y. Na to slúžia **miery kontingencie**, ktoré určujú stupeň závislosti v intervale $\langle 0, 1\rangle$, pričom hodnoty blízke 0 ukazujú na slabú závislosť. Medzi najznámejšie miery kontingecie patria **Pearsonov koeficient kontingecie** a **Cramerov koeficient kontingencie**. Ďalšie miery kontingencie sa nachádzajú napr. v [5].

3.3.1 Pearsonov koeficient kontingencie

$$C_p = \sqrt{\frac{P}{P+n}}. \tag{3.4}$$

Tento koeficient nadobúda hodnoty z intervalu $\langle 0; \sqrt{(q-1)/q}\rangle$, kde $q = min\{a, b\}$. Problémom je, že nemá stále maximum, ktoré závisí od počtu riadkov a stĺpcov kontingenčnej tabuľky. Preto sa pri porovnávaní intenzity závisloti v rôznych tabuľkách využíva podiel $C_p/max\ C_p$.

3.3.2 Cramerov koeficient kontingencie

Tento koeficient známy aj pod názvom **Cramerovo V** má tvar

$$V = \sqrt{\frac{P}{n(q-1)}}, \tag{3.5}$$

kde $q = min\{a, b\}$.

V menovateli je maximálna hodnota, ktorú môže Pearsonova štatistika chí-kvadrát dosiahnuť. Vďaka tejto konštrukcii nadobúda pri úplnej závislosti hodnotu 1. Je považovaný za jeden z najlepších mier kontingencie.

4 Časové rady

Táto kapitola sa venuje časovým radom, ktoré považujem za ideálnu metódu pri skúmaní vývoja počtu požiarov za dané časové obdobie. Tomuto vývoju sa budem venovať v kapitole 7.

Postupnosť hodnôt určitého kvantitatívneho ukazovateľa usporiadaných z hľadiska času v smere od minulosti do prítomnosti nazývame **časový rad**. Je to dvoj a viacrozmerný výber dát, kde závislá premenná y je jednoznačne viazaná na čas. Časové intervaly sú väčšinou ekvidistantné (rovnomerne rozdelené) a vďaka tomu môžeme časový rad zapísať v tvare

$$y_t, t = 1, \ldots, n,$$

kde y_t je náhodná premenná a t je časová premenná s počtom pozorovaní n.

4.1 Delenie časových radov

Časové rady majú z rôznych hľadísk množstvo delení. Uveďme si najzákladnejšie z nich.

Rozdelenie z časového hľadiska

Intervalový časový rad - dáta závisia na dĺžke intervalu, počas ktorého boli sledované. Hodnoty tohto časového radu má zmysel sčítavať, prípadne priemerovať za podmienky, že jednotlivé intervaly sú rovnako dlhé. Ak táto podmienka nie je splnená, treba vykonať tzv. **očistenie časového radu** (prepočítať hodnoty na jednotkový časový interval).

Okamihový časový rad - dáta sa vzťahujú k určitému okamihu. Sčítavanie v tomto prípade nemá zmysel, na priemerovanie sa využívajú chronologické priemery.

Ak sú jednotlivé intervaly rovnako dlhé použijeme **chronologický priemer**

$$\bar{y} = \frac{\frac{y_1}{2} + y_2 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1}. \tag{4.1}$$

V opačnom prípade použijeme **vážený chronologický priemer**

$$\bar{y} = \frac{\frac{y_1+y_2}{2}d_1 + \frac{y_2+y_3}{2}d_2 + \ldots + \frac{y_{n-1}+y_n}{2}d_{n-1}}{d_1 + d_2 + \ldots + d_{n-1}}, \tag{4.2}$$

kde d_i je dĺžka jednotlivých intervalov.

Rozdelenie z hľadiska periodicity

Pod pojmom periodicita rozumieme dĺžku časového intervalu, resp. rozpätie medzi okamihmi. Ak je toto časové rozmedzie kratšie ako jeden rok (štvrťrok, mesiac, deň, ...) ide o **krátkodobý časový rad**. V druhom prípade sa jedná o **dlhodobý časový rad**. V praktickej časti bakalárskej práce (kapitola 7) sa vyskytuje krátkodobý aj dlhodobý časový rad.

4.2 Základné charakteristiky

K posúdeniu vývoja časových radov nám pomôžu rôzne charakteristiky. Okrem úplne najzákladnejších ako sú (vážený) aritmetický priemer, smerodatná odchýlka, rozptyl a (vážený) chronologický priemer existujú ďalšie charakteristiky, tzv. **miery dynamiky**, ktoré charakterizujú vývoj časového radu. Medzi základné patria:

Absolútny prírastok (prvá diferencia)
určuje „o koľko“ sa zmenil časový rad medzi jednotlivými časovými okamihmi.

$$\Delta y_t = y_t - y_{t-1}, \tag{4.3}$$

kde $t = 2, \ldots, n$

Priemerný absolútny prírastok
určuje „o koľko“ sa priemerne zmenil časový rad v období medzi dvoma meraniami.

$$\bar{\Delta} y_t = \frac{y_n - y_1}{n - 1}. \tag{4.4}$$

Koeficient (tempo) rastu
určuje „koľkokrát“ sa zmenil časový rad medzi jednotlivými časovými okamihmi.

$$k_t = \frac{y_t}{y_{t-1}}, \tag{4.5}$$

kde $t = 2, \ldots, n$

Priemerný koeficient rastu
určuje „koľkokrát“ sa priemerne zmenil časový rad v období medzi dvoma meraniami.

$$\bar{k} = \sqrt[n-1]{\frac{y_n}{y_1}}. \tag{4.6}$$

Relatívny prírastok
určuje „o koľko percent“ sa zmenil časový rad medzi jednotlivými okamihmi.

$$\delta_t = \frac{y_t}{y_{t-1} - 1} \cdot 100, \tag{4.7}$$

kde $t = 2, \ldots, n$

Priemerný relatívny prírastok
určuje „o koľko percent“ sa priemerne zmenil časový rad v období medzi dvoma meraniami.

$$\bar{\delta} = (\bar{k} - 1) \cdot 100. \tag{4.8}$$

4.3 Metódy analýzy časových radov

Metódy analýzy časových radov sa snažia nájsť akési pravidelnosti a systematiku v správaní sa časového radu. Čím presnejšie ich určíme a odmeráme, tým vierohodnejšie budú odhady do budúcnosti. Výber správnej metódy závisí od množstva faktorov ako sú napr. účel analýzy, typ časového radu apod. Medzi základné metódy patria:

- Klasický model (dekompozičná metóda)
- Boxova - Jenkinsova metodológia
- Lineárne dynamické modely
- Spektrálna analýza

Tieto metódy sú pomerne rozsiahle a náročné. Ďalej sa budeme venovať len dekompozičnej metóde. Stručný popis k ostatným metódam nájdete napr. v [7].

Dekompozičná metóda

Cieľom tejto metódy je snaha rozložiť časový rad na súčet alebo súčin štyroch základných zložiek. Medzi ne patria:

- **Trendová zložka** (T_t) - dlhodobá tendencia vo vývoji časového radu. Môže mať rastúci, klesajúci, prípadne stacionárny charakter. Zvykne sa modelovať pomocou matematických kriviek.
- **Cyklická zložka** (C_t) - kolísanie okolo trendovej zložky, v ktorom sa striedajú fázy rastu a poklesu. Jednotlivé cykly majú nepravidelný charakter (rôzna dĺžka a amplitúda) a vytvárajú sa za obdobie dlhšie ako 1 rok.
- **Sezónna zložka** (S_t) - pravidelné kolísanie okolo trendovej zložky s periódou maximálne 1 rok.
- **Náhodná zložka** (E_t) - táto zložka nemá rozpoznateľný charakter. Tvoria ju náhodné a nesystematické výkyvy a chyby v meraní.

Podľa spôsobu rozkladu rozlišujeme dva typy modelu:

Aditivný (súčtový) model

$$y_t = T_t + C_t + S_t + E_t$$

Multiplikatívny (súčinový) model

$$y_t = T_t \cdot C_t \cdot S_t \cdot E_t$$

Aditívna dekompozícia sa používa ak variabilita hodnôt časového radu je približne konštantná v čase, ak rastie alebo sa v čase mení, tak sa používa multiplikatívny model. Rozdiel medzi oboma je, že u aditívneho modelu sú všetky zložky časového radu v rovnakých jednotkách ako pôvodný rad, kým u multiplikatívneho modelu je v rovnakých jednotkách iba trendová zložka a ostatné sú vyjadrené relatívne. V praxi sa častejšie využíva aditívny model, naviac multiplikatívny tvar sa dá logaritmovaním previesť na aditívny.

Analýza trendovej zložky

Dlhodobá vývojová tendencia je základným kameňom pre čo najpresnejšie odhady do budúcnosti. Preto presná identifikácia a popis trendovej zložky patria medzi najdôležitejšie úlohy pri modelovaní časových radov. Na získanie trendovej zložky využívame:

- **Jednoduché grafické metódy** ako sú napr. metóda vybalancovania výkyvov, spriemerňovania cyklov a iné.
- **Trendové funkcie**, ktoré vyrovnávajú, vyhladzujú časový rad v jednom kroku.
- **Adaptívne prístupy**, kde patria exponencionálne vyrovnávanie a metódy kĺzavých priemerov, ktoré vyrovnávajú časový rad postupne vo viacerých krokoch.

Ďalej sa budem venovať iba dekompozícii časového radu na odhad počtu požiarov v následujúcich rokoch. Viac informácií nájdete napr. v [8, 9].

4.3.1 Metóda kĺzavých priemerov

Táto metóda patrí medzi adaptívne prístupy, teda dokáže reagovať na zmeny v charaktere trendovej zložky. Názov je odvodený od „kĺzania sa“ pri výpočte priemerov po časovom rade vždy o jedno pozorovanie dopredu. Princíp spočíva v rozdelení časového radu y_t na kratšie časové úseky. Dĺžka úseku je rovná lineárnej kombinácii $2p + 1$ členov pôvodného radu. Pričom prvých a posledných p hodnôt časového radu zostáva nevyrovnaných. Sila vyrovnania závisí od dĺžky kĺzavého priemeru.

Kĺzavé priemery rozdeľujeme podľa ich dĺžky do dvoch skupín.

Jednoduché kĺzavé priemery rádu 2p+1

Využívame ich ak dĺžka kĺzavého priemeru je „nepárna". Časový úsek s dĺžkou $2p+1$ vyrovnáme pomocou aritmetického priemeru.

$$\bar{y_t} = \frac{y_{t-p} + y_{t-p+1} + \ldots + y_{t+p-1} + y_{t+p}}{2p+1}, \tag{4.9}$$

kde $t = p+1, \ldots, n-p$.

Centrované kĺzavé priemery rádu 2p

V opačnom prípade t.j. ak má časový úsek dĺžku $2p$, používame centrované kĺzavé priemery. Nepárny počet členov v kĺzavom priemere získame tak, že namiesto prvého členu vezmeme priemer prvej a poslednej hodnoty.

$$y_t = \frac{1}{4p}(y_{t-p} + 2y_{t-p+1} + \ldots + 2y_{t+p-1} + y_{t+p}), \tag{4.10}$$

kde $t = p+1, \ldots, n-p$.

4.3.2 Očistenie časového radu od sezónnych vplyvov

Sezónne vplyvy komplikujú sledovanie vývoja a zneprehľadňujú časové rady. Na odstránenie sezónnej zložky existuje množstvo metód. Medzi najjednoduchší spôsob patrí očisťovanie časového radu s využitím sezónnych parametrov.

Najprv určíme **sezónny faktor** tak, že odpočítame centrované kĺzavé priemery (kde dĺžka je rovná perióde časovej rady) od skutočných nameraných hodnôt. Potom z týchto rozdielov vypočítame priemernú odchýlku pre každú individuálnu časť periódy a dostaneme výsledný sezónny faktor. Pre lepšiu predstavu si opíšme výpočet sezónneho faktora pre konrétny prípad, ktorý budem používať v kapitole 7.2.4. Keďže analyzujem časový rad s dĺžkou periódy jeden rok, určujem sezónny faktor pre jednotlivé mesiace. Vypočítam odchýlku 12-členných kĺzavých priemerov od pôvodného časového radu. Potom januárový sezónny faktor vypočítam ako priemer všetkých januárových odchýliek a obdobne postupujem u každého mesiaca. Z toho je viditeľné, že sezónny faktor je pre každú periódu rovnaký. Očistený časový rad vznikne odpočítaním sezónneho faktoru od nameraných hodnôt.

4.4 Regresná analýza

Keďže pri skúmaní vývoja požiarov v kapitole 7 bude využívaná regresná priamka, tak si v skratke popíšme, čo je podstatou jednoduchej lineárnej regresie. Rozsiahlejšie vysvetlenie a ostatné druhy regresie sú popísané napr. v [1, 8].

Regresná analýza ako taká skúma závislosť medzi náhodnými veličinami. Jednoduchá znamená, že skúma závislosť vysvetľovanej (závislej) premennej Y a vysvetľujúcej (nezávislej) premennej X (často označovaná ako regresor). Lineárna značí, že hľadáme funkciu, ktorá je lineárna v parametroch (alebo sa dá na takúto funkciu previesť). Funkcia má potom tvar

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \tag{4.11}$$

kde β_0,β_1 sú regresné koeficienty a e_i je náhodná chyba.

Vo väščine prípadov nedokážeme regresné koeficienty určiť presne, preto ich nahradzujeme odhadmi b_0,b_1. Odhad lineárnej funkcie má tvar

$$\hat{Y_i} = b_0 + b_1 X_i. \tag{4.12}$$

Regresné koeficienty najčastejšie odhadujeme pomocou metódy najmenších štvorcov (chceme aby súčet druhých mocnín chýb riešenia bol čo najmenší). Označme si chybu riešenia (reziduum) ako $e_i = Y_i - \hat{Y_i}$. Potom súčet štvorcov reziduí, ktorý chceme minimalizovať má po krátkej úprave tvar

$$\phi = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$

Ak chceme použiť metódu najmenších štvorcov, musia byť splnené predpoklady lineárneho regresného modelu, ktoré sú uvedené napr. v [6]. Potom po jednoduchých úpravách (viď napr. [8]) dostaneme odhady regresných koeficientov.

$$b_0 = \frac{\sum\limits_{i=1}^{n} Y_i}{n} - b_1 \frac{\sum\limits_{i=1}^{n} X_i}{n}, \qquad b_1 = \frac{n \sum\limits_{i=1}^{n} Y_i X_i - \sum\limits_{i=1}^{n} Y_i \sum\limits_{i=1}^{n} X_i}{n \sum\limits_{i=1}^{n} (X_i)^2 - \Big(\sum\limits_{i=1}^{n} X_i \Big)^2}.$$

Keď už máme vytvorený regresný model, ostáva nám ho overiť (verifikovať). Jedná sa napríklad o overenie kvality a stability modelu, predpokladov použitia metódy najmenších štvorcov a iné. V ďalšom texte sa budeme venovať len overeniu stability modelu pomocou t-testu, ostané verifikácie modelu sú uvedené napr. v [1].

Pomocou t-testu overujeme, či nemôžeme z modelu vypustiť jednotlivé regresory, prípadne konštantu, teda nulovosť koeficientov β_0 a β_1. Testujeme nulovú hypotézu H_0 : $\beta_i = 0$ oproti aleternatíve $H_A : \beta_i \neq 0$.
Testová štatistika má tvar

$$S = \frac{b_i - \beta_i}{s_{b_i}}, \tag{4.13}$$

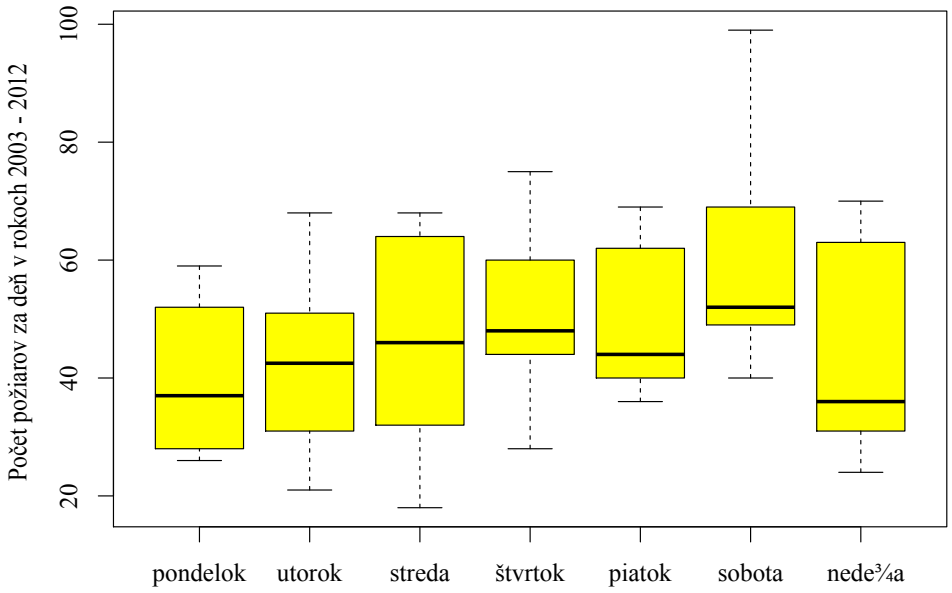
kde s_{b_i} je smerodatná odchýlka odhadu.

P-hodnotu vypočítame podľa vzťahu $2\ min\{F_0(x_{OBS}; 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribučná funkcia Studentovho rozdelenia s $n - k + 1$ stupňami voľnosti. V prípade nezamietnutia nulovej hypotézy môžeme príslušný regresor z modelu odstrániť.

5 Počet požiarov v priebehu týždňa

V tejto kapitole sa budem snažiť pomocou analýzy rozptylu skúmať počet požiarov v jednotlivých dňoch týždňa medzi rokmi 2003 – 2012. V úvode bakalárskej práce boli uvedené podrobnejšie informácie o OR HaZZ v Spišskej Novej Vsi.

Pre lepšiu predstavu a prehľadnosť si počet požiarov v jednotlivých dňoch týždňa zobrazme na obrázku 5.1. V tabuľke 5.1 budú uvedené základné súhrnné štatistiky.



Obrázok 5.1: Počet požiarov v jednotlivých dňoch týždňa v rokoch 2003–2012

	Pondelok	Utorok	Streda	Štvrtok	Piatok	Sobota	Nedeľa
$\bar{x}$	40,2	43,0	46,9	51,5	49,1	59,8	44,5
min	26,0	21,0	18,0	28,0	36,0	40,0	24,0
$x_{0,25}$	29,5	32,0	33,8	44,5	40,5	49,3	32,0
$x_{0,5}$	37,0	42,5	46,0	48,0	44,0	52,0	36,0
$x_{0,75}$	50,0	50,8	62,5	58,3	59,5	69,0	60,3
max	59,0	68,0	68,0	75,0	69,0	99,0	70,0
IQR	20,5	18,8	28,8	13,8	19,0	19,8	28,3
s^2	156,4	227,6	286,3	209,6	131,7	327,3	312,3
s	12,5	15,1	16,9	14,5	11,5	18,1	17,7

Tabuľka 5.1: Súhrnné štatistiky pre počet požiarov v priebehu týždňa

Z obrázka 5.1 a tabuľky 5.1 vidíme, že najvyšší počet požiarov je vo štvrtky a soboty. Napríklad v sobotu maximum nadobúda hodnoty takmer 100, čo odpovedá skutočnosti, že v roku 2012 bolo za všetkých 52 sobôt ohlásených 99 požiarov (priemerný počet požiarov na deň týždňa počas roka je približne 48). Z hodnoty mediánu vidíme, že v polovici všetkých sobôt bol počet požiarov menší ako 52. Vplyv extrémne vysokého počtu požiarov v soboty roku 2012 sa odráža aj na výberovom smerodatnom rozptyle (resp. výberovej smerodatnej odchýlke).

Na základe polohy mediánu a „výšky krabičky“ (IQR) predpokladám, že rozdiel v počte požiarov za jednotlivé dni týždňa nie je štatisticky významný. Túto vyslovenú domnienku sa budem v nasledujúcej časti snažiť potvrdiť alebo vyvrátiť.

Najprv si zhrňme postup pri analýze rozptylu a rozdeľme ho do následujúcich základných krokov:

1. Overenie predpokladov testu
2. Nulová a alternatívna hypotéza
3. Výpočet p-hodnoty
4. Post hoc analýza

5.1 Overenie predpokladov testu

Ako už bolo spomenuté v predošlej podkapitole 2.1.1, aby sme mohli začať samotné testovanie, musia byť splnené predpoklady normality a homoskedasticity dát. Na overenie normality dát použijem Shapiro–Wilkov test a v závislosti na výsledku na overenie homoskedasticity použijem Bartlettov alebo Leveneov test. Oba testy vykonám pomocou vstavaných funkcií v štatistickom softvére „R“.

Normalita dát

Testujeme nulovú hypotézu H_0 : Výber pochádza z normálneho rozdelenia, oproti H_A : Výber nepochádza z normálneho rozdelenia.

Pripomeňme, že normalitu dát overujeme samostatne pre každý výber – deň v týždni. Kvôli prehľadnosti, zobrazme výsledky testu v tabuľke 5.2.

Deň v týždni		p-hodnota Shapirovho-Wilkovho testu
pondelok	(1)	0,195
utorok	(2)	0,920
streda	(3)	0,556
štvrtok	(4)	0,593
piatok	(5)	0,127
sobota	(6)	0,141
nedeľa	(7)	0,082

Tabuľka 5.2: Overenie normality dát (počet požiarov v priebehu týždňa)

Z vyššie uvedenej tabuľky 5.2 vidíme, že p-hodnota v žiadnom z prípadov neklesla pod hodnotu 0,05. Nulové hypotézy na hladine významnosti 0,05 nezamietame a teda predpokladáme, že výbery pochádzajú z normálneho rozdelenia.

Homoskedasticita

Vzhľadom k tomu, že výbery (počty požiarov) pochádzajú z normálneho rozdelenia, homoskedasticita bude overená pomocou Bartlettovho testu.

Testujeme nulovú hypotézu $H_0 : {\sigma_1}^2 = {\sigma_2}^2 = \ldots = {\sigma_7}^2$, teda, že rozptyly sa v jednotlivých dňoch štatisticky významne nelíšia, oproti alternatíve, že rozptyly sa aspoň v 2 dňoch štatisticky významne líšia.

Test	p-hodnota
Bartlettov test	0,811

Tabuľka 5.3: Overenie homoskedasticity dát

Tak ako aj v predošlom prípade, ani teraz nulovú hypotézu na hladine významnosti 0,05 nezamietame a predpoklad homoskedasticity dát považujeme za overený.

5.2 Nulová a alternatívna hypotéza

Na začiatok stanovíme nulovú a alternatívnu hypotézu.

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_7$, teda priemerný počet požiarov sa medzi jednotlivým pozorovanými dňami štatisticky významne nelíši.

$H_A : \neg H_0$, teda priemerný počet požiarov sa v aspoň 2 dňoch štastisticky významne líši.

5.3 Výpočet p-hodnoty

Ďalej vykonáme samotný výpočet. Kedže softvér „R“ nám nevypočíta celú tabuľku ANOVA, doplníme ju ručne podľa tabuľky 2.4. Výsledná tabuľka vyzerá následovne:

Zdroj premenlivosti	Variabilita	Stupne voľnosti	Rozptyl	F-pomer	p-hodnota
Medzitriedny	2518,6	6	419,8	1,78	0,12
Vnútorný	14860,0	63	235,9		
Celkový	17378,6	69			

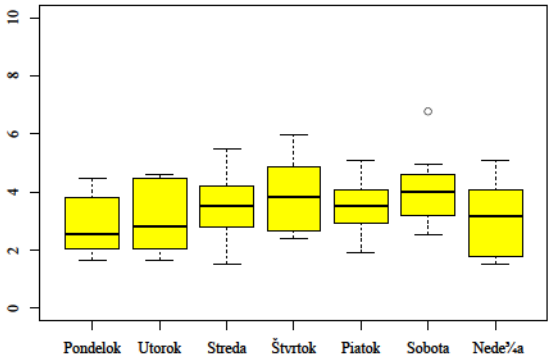
Tabuľka 5.4: Doplnená tabuľka ANOVA

Na základe p-hodnoty uvedenej v tabuľke 5.4 nulovú hypotézu nezamietame, teda môžeme tvrdiť, že počet požiarov sa v jednotlivých dňoch týždňa štatisticky významne nelíši. Teda domnienka vyslovená na základe obrázku 5.1 bola pravdivá.

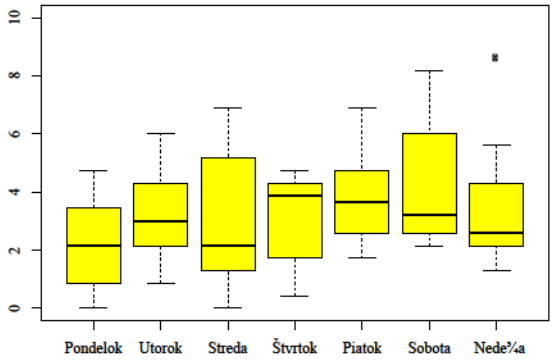
Keďže nulovú hypotézu nezamietame, nemá zmysel zaoberať sa post hoc analýzou.

5.4 Porovnanie počtu požiarov počas týždňa v jednolivých zásahových obvodoch

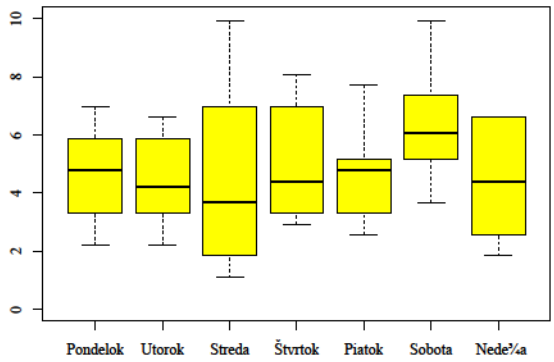
V tejto časti budem porovnávať počty požiarov počas týždňa medzi zásahovými obvodmi Spišská Nová Ves, Gelnica a Krompachy. Postup pri výpočte bude totožný s predchádzajúcim, takže jednotlivé popisy budú omnoho stručnejšie. Počet požiarov v tomto prípade bude prerátaný na 10000 obyvateľov.



(a) Spišská Nová Ves



(b) Gelnica



(c) Krompachy

Obrázok 5.2: Počet požiarov počas týždňa v zásahových obvodoch Spišská Nová Ves, Gelnica a Krompachy

Z obrázku 5.2 a tabuliek A.1, A.2, A.3 (viď Príloha A) vidíme, že jednoznačne najvyšší počet požiarov vznikol v obciach spadajúcich pod zásahový obvod Krompachy. Oproti ostatným zásahovým obvodom sú počty požiarov na 10000 obyvateľov vyššie približne o 40%. To môže byť spôsobené napríklad tým, že v zásahovom obvode Krompachy sa nachádza vyšší počet osád, v okolí ktorých je požiarovosť podstatne vyššia. Aj v zásahových obvodoch samostatne platí, že najvyšší počet požiarov je v sobotu. V zásahových obvodoch Spišská Nová Ves a Gelnica je počet požiarov približne rovnaký, rozdiel je len v ich rozptýlenosti. Vidíme, že v Gelnici je táto rozptýlenosť vyššia, čo potvrdzujú aj hodnoty výberovej smerodatnej odchýlky.

Ďalším krokom je overenie normality a homoskedasticity dát v jednotlivch zásahových obvodoch a na základe výsledkov voľba vhodného testu. U dát v zásahových obvodoch Spišská Nová Ves a Krompachy bola pomocou Shapirovho-Wilkovho testu overená normalita dát a pomocou Bartlettovho testu aj homoskedasticita dát. Problém nastal pri dátach pre zásahový obvod Gelnica, kde nebola splnená podmienka normality dát. Z

tohto dôvodu bola homoskedasticita overená pomocou Leveneovho testu. Na záver sme overovali platnoť nulovej alebo alternatívnej hypotézy, ktorých tvar je rovnaký ako v predošlom prípade (5.2). V následujúcej tabuľke 5.5 sú zobrazené príslušné testy a ich p-hodnoty pre jednotlivé zásahové obvody.

Zásahový obvod	**Test**	**p-hodnota**
Spišská Nová Ves	ANOVA	0,149
Gelnica	Kruskalov-Wallisov test	0,430
Krompachy	ANOVA	0,205

Tabuľka 5.5: Test a p-hodnota pre jednotlivé zásahové obvody na určenie závislosti počtu požiarov na dni týždňa

Na základe p-hodnôt v tabuľke 5.5 nulové hypotézy nezamietame, čo znamená, že v ani jednom zásahovom obvode sa počet požiarov počas týždňa štatisticky významne nelíši. Keďže sme nulové hypotézy nezamietli, nebudeme sa zaoberať post hoc analýzou.

6 Závislosť príčiny požiaru na zásahovom obvode

V tejto časti mojej bakalárskej práce budem pomocou kontingenčných tabuliek skúmať, či príčina požiaru závisí na zásahovom obvode, v ktorom požiar vznikol. Jedná sa o hasičské stanice v obciach Spišská Nová Ves, Gelnica a Krompachy. Príčiny požiarov sú rozdelené do ôsmich skupín, ktoré odpovedajú oficiálnemu číselníku príčin Ministerstva vnútra Slovenskej republiky.

6.1 Kontingenčná tabuľka

Čo konkrétne si predstavujeme pod pojmom kontingenčná tabuľka sme si opísali v kapitole 3.1. Uveďme si teda kontingenčnú tabuľku závislosti príčiny požiaru na zásahovom obvode. Príčiny požiarov sú kvôli prehľadnosti tabuľky očíslované. Názvy, ku ktorým sú priradené číslené hodnoty nájdete v Príloha B.

	1	2	3	4	5	6	7	8	Spolu
Gelnica	16	2	382	20	42	50	6	21	539
Krompachy	13	4	800	12	37	38	2	19	925
Spišská Nová Ves	16	16	1563	23	73	131	7	57	1886
Spolu	45	22	2745	55	152	219	15	97	3350

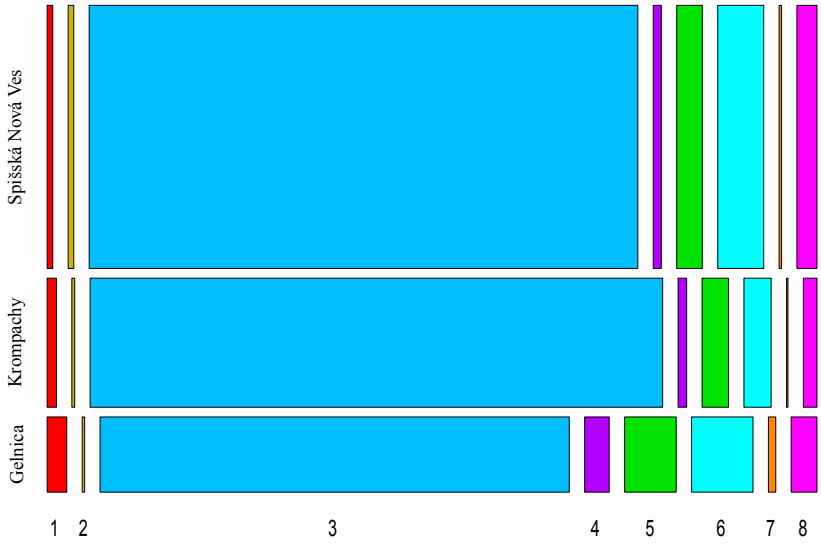
Tabuľka 6.1: Kontingenčná tabuľka závislosti príčiny požiaru na zásahovom obvode

V tabuľke sú uvedené absolútne, marginálne početnosti a celkový rozsah výberu. Ako bolo uvedené v kapitole 3.1, tabuľku môžeme rozšíriť o relatívne, riadkové relatívne a stĺpcové relatívne početnosti.

	1	2	3	4	5	6	7	8	Spolu
Gelnica	16 0,005 0,030 0,356	2 0,001 0,004 0,091	382 0,114 0,709 0,139	20 0,006 0,037 0,364	42 0,013 0,078 0,276	50 0,015 0,093 0,228	6 0,002 0,011 0,400	21 0,006 0,039 0,216	539
Krompachy	13 0,004 0,014 0,289	4 0,001 0,004 0,182	800 0,239 0,865 0,291	12 0,004 0,013 0,218	37 0,011 0,040 0,243	38 0,011 0,041 0,174	2 0,001 0,002 0,133	19 0,006 0,021 0,196	925
Spišská Nová Ves	16 0,005 0,008 0,356	16 0,005 0,008 0,727	1563 0,467 0,829 0,569	23 0,007 0,012 0,418	73 0,022 0,039 0,480	131 0,039 0,069 0,598	7 0,002 0,004 0,467	57 0,006 0,030 0,588	1886
Spolu	45	22	2745	55	152	219	15	97	3350

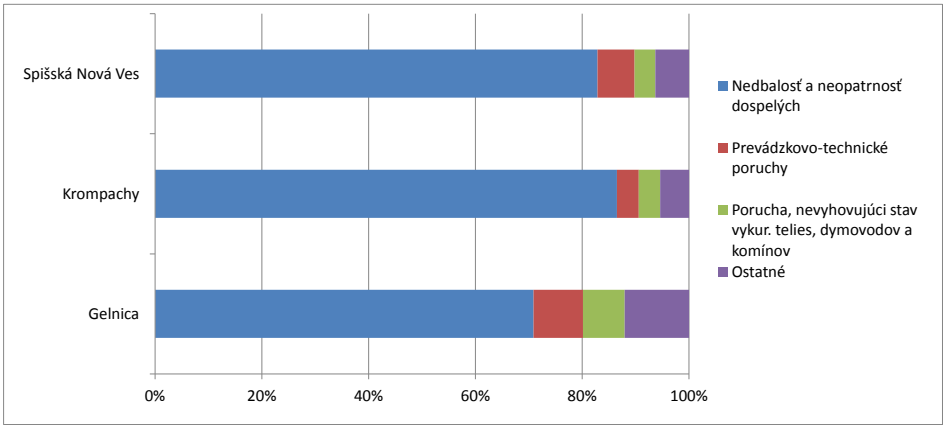
Tabuľka 6.2: Rozšírená kontingenčná tabuľka závislosti príčiny požiaru na zásahovom obvode (relatívne, riadkové relatívne a stĺpcové relatívne početnosti)

Grafickou obdobou kontingenčnej tabuľky je **mozaikový graf**. Ten pozostáva z a rád obdĺžnikov, pričom každá z nich obsahuje b obdĺžnikov. Dĺžky strán sú úmerné príslušným marginálnym relatívnym početnostiam. Ak by bol graf tvorený zvislými „pruhmi“, vypovedalo by to o nezávisosti sledovaných znakov. Naopak, čím zložitejší je mozaikový graf, tým je závislosť väčšia.



Obrázok 6.1: Mozaikový graf

Okrem mozaikového grafu sa používajú aj **100% skladaný pruhový graf**, ktorý neberie v úvahu riadkové marginálne početnosti, **zhlukový** alebo **kumulatívny stĺpcový graf**.



Obrázok 6.2: 100% skladaný pruhový graf

Na obrázku 6.1 vidíme mozaikový graf zobrazujúci rozloženie príčin požiarov v jednotlivých zásahových obvodoch. Jednoznačne najrozšírenejšou príčinou vo všetkých zásahových obvodoch je príčina číslo 3, teda nedbalosť a neopatrnosť dospelých. Okrem toho sa najčastejšie vyskytujú príčiny odpovedajúce číslam 5 (porucha, nevyhovujúci stav vykurovacích telies, dymovod a komínov) a 6 (prevádzkovo-technické poruchy). Podľa členitosti mozaikového grafu môžeme na prvý pohľad odhadnúť, že príčina požiaru závisí na zásahovom obvode. Tento odhad overíme v nasledujúcej podkapitole 6.2.

100% skladaný pruhový graf na obrázku 6.2 znázorňuje výskyt troch najčastejších príčin požiarov, zvyšné príčiny sú zahrnuté v skupine ostatné. Je viditeľné, že nedbalosť a neopatrnosť dospelých zaberá približne 70% - 85% všetkých príčin. Príčiny číslo 5 a 6 odpovedajú približne 5% - 10%. Ďalšie príčiny sú takmer zanedbateľné.

6.2 Testovanie nezávislosti

Použitie χ^2 testu o nezávislosti alebo Yatesovej korekcie χ^2 testu o nezávislosti závisí od splnenia podmienok testu uvedených v kapitole 3.2.1. K ich overeniu potrebujeme poznať očakávané početnosti m_{ij}, ktoré si pre prehľadnosť uveďme v tabuľke 6.3.

	1	2	3	4	5	6	7	8	Spolu
Gelnica	16 7,240	2 3,540	382 441,658	20 8,849	42 24,456	50 35,236	6 2,413	21 15,607	539
Krompachy	13 12,425	4 6,075	800 757,948	12 15,187	37 41,970	38 60,470	2 4,142	19 26,784	925
Sp. N. Ves	16 25,334	16 12,386	1563 1545,394	23 30,964	73 85,574	131 123,294	7 8,445	57 54,610	1886
Spolu	45	22	2745	55	152	219	15	97	3350

Tabuľka 6.3: Kontingenčná tabuľka závislosti príčiny požiaru na zásahovom obvode (rozšírená o očakávané početnosti)

Z vyššie uvedenej tabuľky 6.3 vidíme, že žiadna z očakávaných početností neklesla pod 2 a že aspoň 80% (20) očakávaných početností je väčších ako 5. Keďže predpoklady tesu považujeme sa splnené, môžeme sformulovať nulovú a alternatívnu hypotézu, ktoré majú tvar:

H_0 : Počet požiarov podľa príčiny sa v jednotlivých zásahových obvodoch štatisticky významne **nelíši**.

H_A : Počet požiarov podľa príčiny sa v jednotlivých zásahových obvodoch štatisticky významne **líši**.

Pomocou softvéru „R“ vypočítame pozorovanú hodnotu Pearsonovej štatistiky chí-kvadrát a príslušnú p-hodnotu.

Testová štatistika	Pozorovaná hodnota	Stupne voľnosti	p-hodnota
Pearsonov chí-kvadrát	84,81	14	$\ll 0{,}001$

Tabuľka 6.4: Pearsonova štatistika chí-kvadrát

Keďže p-hodnota $\ll 0{,}001$, na hladine významnosti 0,05 zamietame nulovú hypotézu v prospech alternatívnej. Odhad vyslovený na základe mozaikového grafu bol správny a môžeme tvrdiť, že počet požiarov podľa príčiny sa v jednotlivých zásahových obvodoch líši.

6.3 Miery závislosti

V kapitole 3.3 boli opísané najčastejšie používané miery kontingencie.

Pearsonov koeficient kontingencie

Podľa vzorca 3.4 vypočítame:

$$C_p = \sqrt{\frac{84{,}81}{84{,}81+3350}} = 0,16.$$

Tento koeficient nadobúda hodnoty z intervalu $\langle 0; 0,82\rangle$, normovaná hodnota je

$$C_p = \frac{0{,}68}{0{,}82} = 0,20.$$

Cramerovo V

Podľa vzorca 3.5 vypočítame:

$$V = \sqrt{\frac{84{,}81}{3350(3-1)}} = 0,11.$$

V predošlej kapitole 6.2 sme zistili, že závislosť medzi príčinou požiaru a zásahovým obvodom je štatisticky významná, no hodnoty koeficientov kontingencie napovedajú, že táto závislosť je pomerne slabá.

7 Vývoj požiarov

V poslednej kapitole budem pomocou časových radov skúmať vývoj požiarov. Konkrétne ako sa vyvýja počet požiarov počas týždňa v jednotlivých mesiacoch a ročný vývoj požiarov v období medzi rokmi 2003 - 2012.

7.1 Vývoj požiarov počas týždňa v jednotlivých mesiacoch

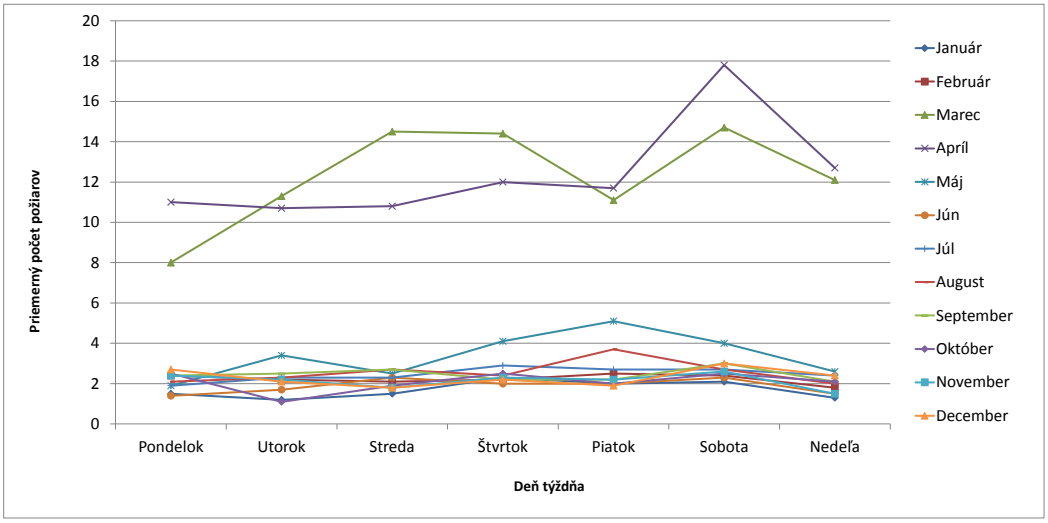
7.1.1 Grafické zobrazenie

Na úvod si uveďme tabuľku, pomocou ktorej vytvoríme graf, vypočítame miery dynamiky a určíme základné vlastnosti časových radov. Hodnoty v tabuľke odpovedajú priemernému počtu požiarov za jeden rok v daný deň týždňa a jednotlivý mesiac.

	január	február	marec	apríl	máj	jún	júl	august	september	október	november	december
pondelok	1,5	2,4	8,0	11,0	1,9	1,4	1,9	2,1	2,4	2,5	2,4	2,7
utorok	1,2	2,2	11,3	10,7	3,4	1,7	2,3	2,3	2,5	1,1	2,2	2,1
streda	1,5	2,1	14,5	10,8	2,5	2,3	2,3	2,7	2,7	1,9	1,8	1,8
štvrtok	2,3	2,2	14,4	12,0	4,1	2,0	2,9	2,4	2,2	2,5	2,3	2,2
piatok	2,0	2,5	11,1	11,7	5,1	2,0	2,7	3,7	2,2	2,0	2,2	1,9
sobota	2,1	2,4	14,7	17,8	4,0	2,3	2,7	2,7	3,0	2,5	2,6	3,0
nedeľa	1,3	1,8	12,1	12,7	2,6	1,5	2,4	2,0	2,1	2,1	1,5	2,4

Tabuľka 7.1: Priemerný počet požiarov počas týždňa v jednotlivých mesiacoch za jeden rok

Vyššie uvedené dáta sú zobrazené v grafe na obrázku 7.1. Z tohto grafu vidíme, že jednoznačne najvyšší počet požiarov je počas víkendu v sobotu. Relatívne vysoká požiarovosť je aj vo štvrtok, prípadne piatok. Priemerný počet požiarov sa vo väčšine mesiacov pohybuje medzi hodnotami 1,5 - 2,5. Výnimkou sú jarné mesiace, najmä marec a apríl, kde je priemerný počet požiarov viditeľne vyšší a nadobúda hodnoty v rozmedzí 8 - 18. Príčinou vysokého počtu požiarov v sobotu môže byť napríklad vo všeobecnosti viacej voľného času u ľudí a s tým spojené výlety v prírode, cestovanie, rôzne opekačky alebo neopatrnosť v dôsledku únavy z pracovného týždňa.



Obrázok 7.1: Vývoj požiarov počas týždňa v jednotlivých mesiacoch

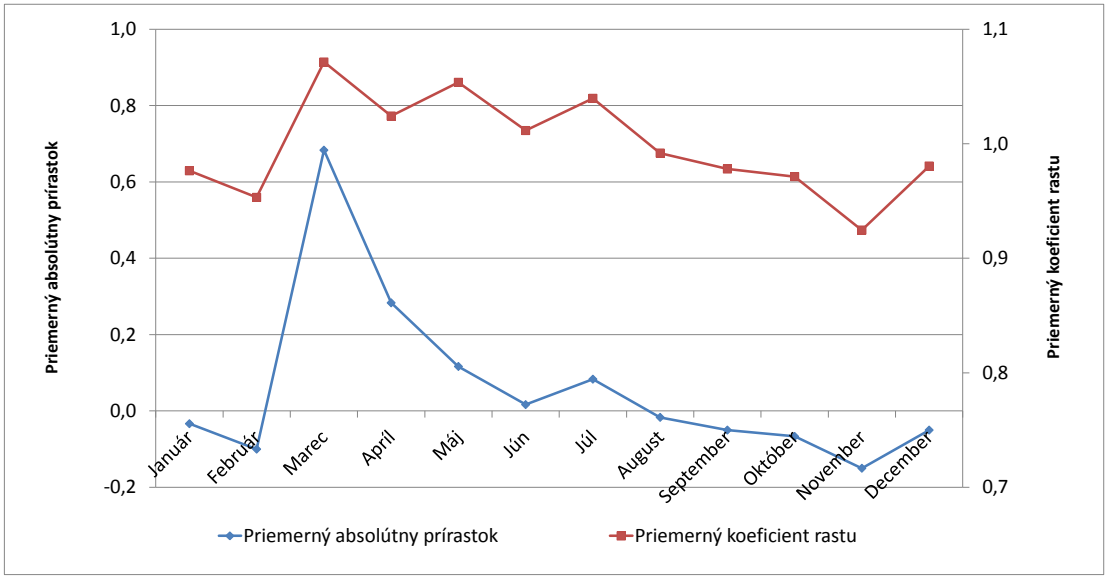
7.1.2 Miery dynamiky

Ako už bolo spomenuté v kapitole 4.2 základné charakteristiky nám umožnia lepšie posúdiť a odhadnúť časový rad. Vybrané miery dynamiky sú zobrazené v nasledujúcej tabuľke a ich grafické zobrazenie na obrázkoch 7.2 a 7.3.

	január	február	marec	apríl	máj	jún	júl	august	september	október	november	december
$\bar{\Delta}$	-0,03	-0,10	0,68	0,28	0,12	0,02	0,08	-0,02	-0,05	-0,07	-0,15	-0,05
$\bar{k}$	0,98	0,95	1,07	1,02	1,05	1,01	1,04	0,99	0,98	0,97	0,92	0,98
$\bar{\delta}$	-2,36	-4,68	7,14	2,42	5,37	1,16	3,97	-0,81	-2,20	-2,86	-7,53	-1,94

Tabuľka 7.2: Miery dynamiky pre vývoj požiarov počas týždňa v jednotlivých mesiacoch

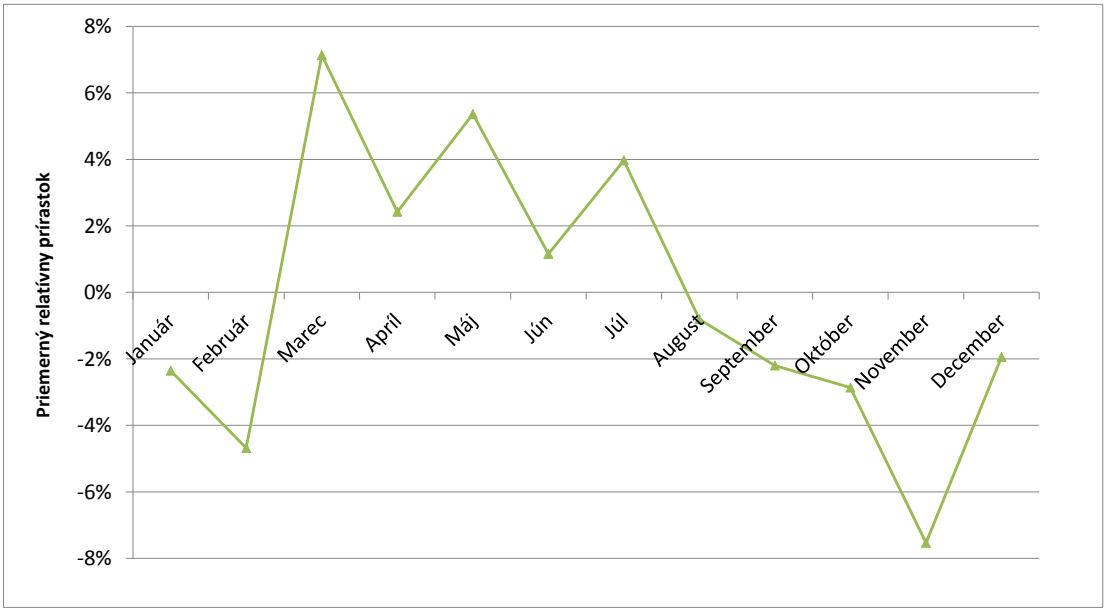
Na obrázku 7.2 vidíme, že primerný absolútny prírastok kolíše okolo hodnoty nula, okrem spomínaných mesiacov marec a apríl, kde je výkyv znateľne vyšší. Kladné hodnoty znamenajú nárast a záporné pokles priemerného počtu požiarov v priebehu týždňa a hovoria nám o koľko priemerne sa zmenil počet požiarov oproti predchádzajúcemu dňu v priebehu týždňa v danom mesiaci. Napríklad v mesiaci marec počet požiarov v priebehu týždňa stúpol o 0,7. Tento obrázok znázorňuje aj priemerný koeficient rastu, ktorého hodnoty sa pohybujú okolo jednotky. Význam týchto hodnôt je rovnaký ako v predošlom prípade, s tým rozdielom, že sa jedná o koľkokrát sa priemerne zmenil počet požiarov v priebehu týždňa oproti predchádzajúcemu dňu v jednotlivých mesiacoch.



Obrázok 7.2: Priemerný absolútny prírastok a priemerný koeficient rastu počtu požiarov počas týždňa v závislosti na mesiaci

Na obrázku 7.3 je znázornený priemerný relatívny prírastok, ktorý na rozdiel od predošlých dvoch hodnôt hovorí, o koľko percent sa zmenil priemerný počet požiarov v priebehu týždňa oproti predošlému dňu v jednotlivých mesiacoch. Napríklad nárast medzi dňami v mesiaci marec bol približne 7%.

Zo všetkých vypočítaných hodnôt môžeme konštatovať, že počet požiarov počas týždňa prudko stúpa v mesiaci marec, následne postupne klesá a minimum nadobúda v zimných mesiacoch.



Obrázok 7.3: Priemerný realtívny prírastok počtu požiarov počas týždňa v závislosti na mesiaci

Keďže kĺzavé priemery a očisťovanie časového radu od sezónnych vplyvov by bolo z dôvodu veľkého počtu (3653 dní) zdĺhavé a náročné na zobrazenie, nebudem sa tomu v tejto časti ďalej venovať.

7.2 Ročný vývoj požiarov v rokoch 2003 - 2012

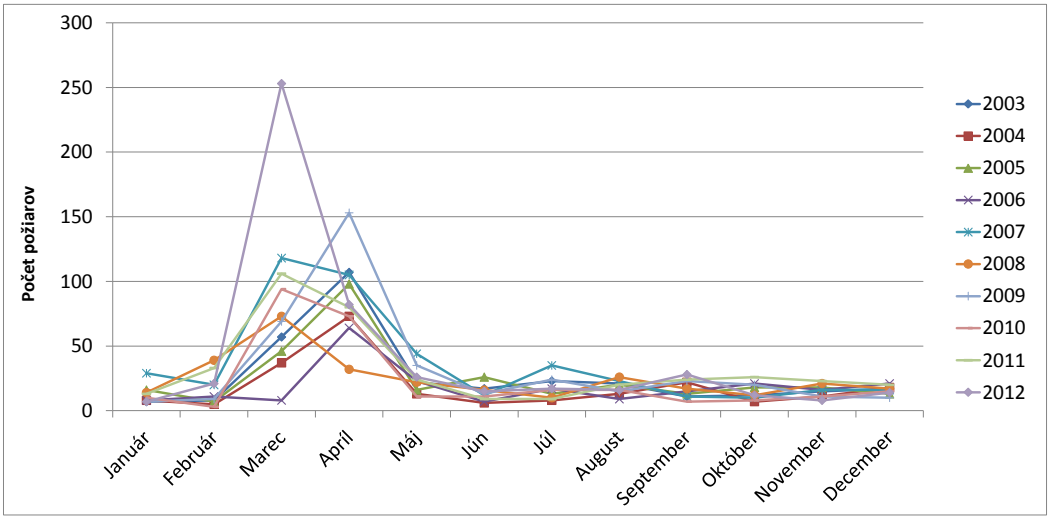
V tejto podkapitole sa budem venovať ročnému vývoju požiarov. Postup pri skúmaní časových radov bude väčšinou rovnaký ako v predošlom prípade. Z tohto dôvodu budú dané časti popísané stručnejšie.

7.2.1 Grafické zobrazenie

Základná tabuľka má nižšie uvedený tvar a jej grafická podoba je na obrázku 7.4.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
január	7	8	16	8	29	14	7	10	13	7
február	8	5	7	11	20	39	9	3	33	21
marec	57	37	46	8	118	73	69	94	106	253
apríl	107	73	98	64	105	32	153	73	80	82
máj	22	13	16	23	44	22	35	11	24	26
jún	17	6	26	7	12	16	13	11	9	15
júl	23	8	13	17	35	10	24	16	9	17
august	21	13	20	9	2	26	15	16	20	16
september	11	22	13	15	11	17	23	7	24	28
október	12	7	18	21	10	12	20	8	26	12
november	15	11	18	16	16	21	11	11	23	8
december	17	18	13	21	16	17	10	15	20	14

Tabuľka 7.3: Počet požiarov v jednotlivých mesiacoch rokov 2003 - 2012



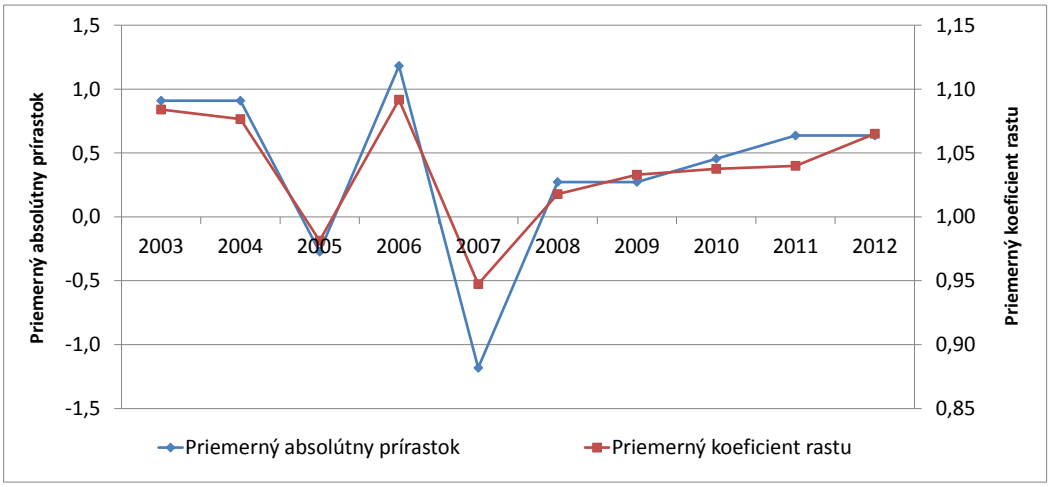
Obrázok 7.4: Ročný vývoj požiarov v rokoch 2003 - 2012

Z obrázka 7.4 je jasne viditeľné, že najvyšší počet požiarov je v mesiacoch marec a apríl. Hodnoty sa pohybujú v rozmedzí 50 - 150 požiarov, okrem marca roku 2012, kde počet požiarov presiahol hodnotu 250. Čo potvrdzuje aj to, že jar spomínaného roka bola výnimočne suchá. V ostatných mesiacoch počet požiarov nepresahuje hodnotu 30. Vývoj požiarov sa za posledných 10 rokov výrazne nezmenil, takže môžeme predpokladať rovnakú tendenciu aj do budúcich rokov.

7.2.2 Miery dynamiky

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
$\bar{\Delta}$	0,91	0,91	-0,27	1,18	-1,18	0,27	0,27	0,45	0,64	0,64
$\bar{k}$	1,08	1,08	0,98	1,09	0,95	1,02	1,03	1,04	1,04	1,07
$\bar{\delta}$	8,40	7,65	-1,87	9,17	-5,26	1,78	3,30	3,75	3,99	6,50

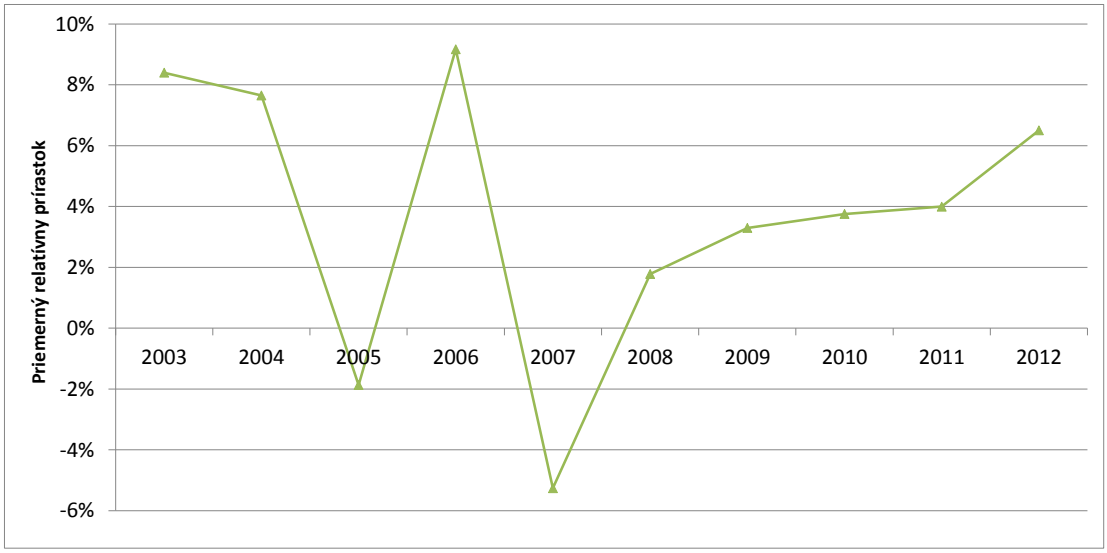
Tabuľka 7.4: Miery dynamiky pre ročný vývoj požiarov v jednotlivých mesiacoch rokov 2003 - 2012



Obrázok 7.5: Priemerný absolútny prírastok a priemerný koeficient rastu ročného vývoja požiarov

Na obrázku 7.5 vidíme, že priemerný absolútny prírastok je medzi rokmi 2005 - 2007 značne rozkolísaný. Vypovedá to napríklad o tom, že počas roka 2006 bol priemerný nárast požiarov o 1,2. V ostatných rokoch bol nárast miernejší, okrem rokov 2005 a 2007 kde išlo o pokles v počte požiarov. Takmer identický priebeh má aj priemerný koeficient rastu, ktorý nadobúda hodnôt 0,95 - 1,1. Tento koeficient nám naznačuje rovnaké správanie časového radu ako priemerný absolútny prírastok.

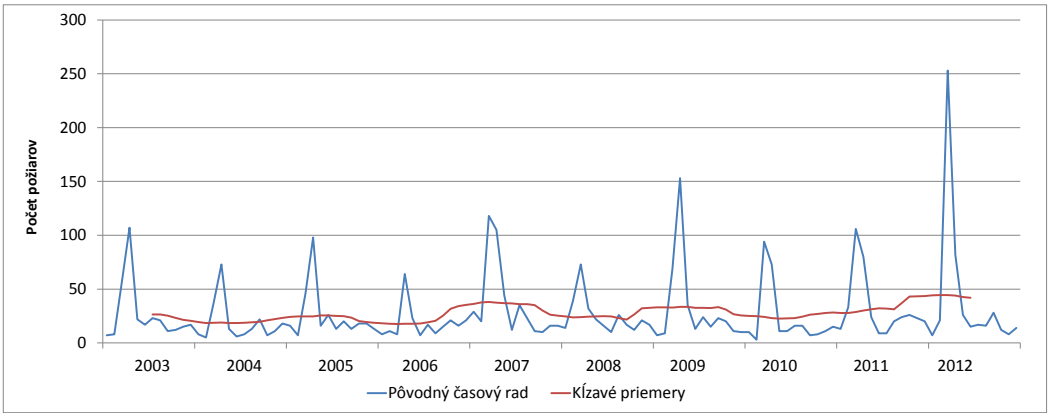
Obrázok 7.6 nám zobrazuje priemerný relatívny prírastok v ročnom vývoji požiarov. Ide o percentuálny nárast resp. pokles počtu požiarov v jednotlivých mesiacoch oproti predošlým. Aj v tomto grafe vidíme výraznejší pokles v rokoch 2005, 2007 a takmer 10% nárast v roku 2006.



Obrázok 7.6: Priemerný relatívny prírastok ročného vývoja požiarov

7.2.3 Kĺzavé priemery

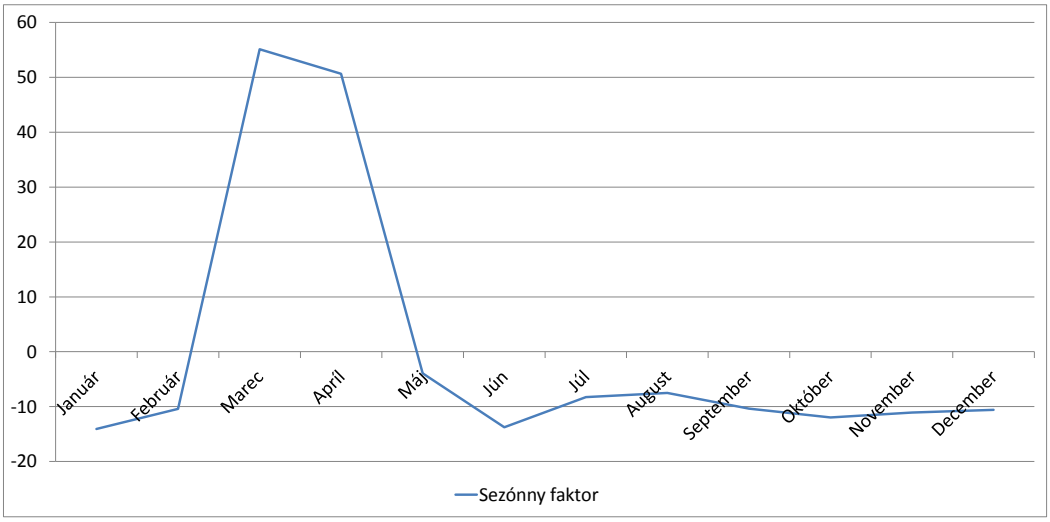
V tomto prípade som pri odhade trendovej zložky zvolila 12-členné kĺzavé priemery, ktoré budem využívať aj pri očisťovaní časového radu od sezónnych vplyvov. Na obrázku 7.7 je viditeľné pomerne silné vyhladenie časového radu. Taktiež môžeme na prvý pohľad odhadnúť, že počet požiarov má jemne stúpajúci charakter.



Obrázok 7.7: Vyhladenie časového radu pomocou 12-členných kĺzavých priemerov

7.2.4 Očistenie časového radu od sezónnych vplyvov

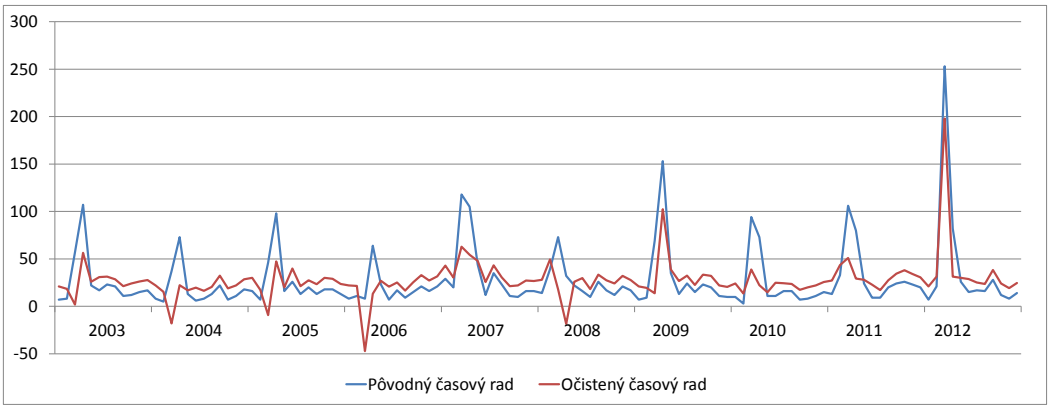
Keďže dĺžka periódy je v tomto prípade 12, sezónny faktor sme pre každý mesiac vypočítali ako priemer príslušných odchýliek centrovaných kĺzavých priemerov dĺžky 12 od pôvodného časového radu. Na obrázku 7.8 je zobrazený sezónny faktor pre jeden rok. Vidíme, že najviac je ovplyvnený mesiacmi marec a apríl, kedy je požiarovosť najvyššia.



Obrázok 7.8: Sezónny faktor

Na obrázku 7.9 je porovnaný pôvodný časový rad s očisteným časovým radom. Je viditeľné, že hodnoty očisteného časového radu sú menej rozkolísané, ba dokonca klesajú pod nulu. To je spôsobené tým, že v niektorých rokoch je počet požiarov v mesiacoch marec a apríl nižší ako hodnota sezónneho faktoru, ktorá v týchto mesiacoch presahuje hodnotu 50. Z tohto dôvodu sa pri rozdiele dostávame do záporných čísel. Dokonca v marci roku 2006 bolo len 8 požiarov a tým pádom sa po odpočítaní sezónneho faktoru dostávame na hodnotu približne -47.

Odhad trendovej zložky pomocou centrovaných kĺzavých priemerov je len predbežný odhad, keďže tento časový rad je skrátený o prvých a posledných 6 hodnôt. Preto je výhodnejšie trendovú zložku predpovedať zo sezónne očisteného radu, na ktorý aplikujeme vhodný model a odhadneme jeho parametre. V tomto prípade budem parametre odhadovať pomocou regresnej analýzy.

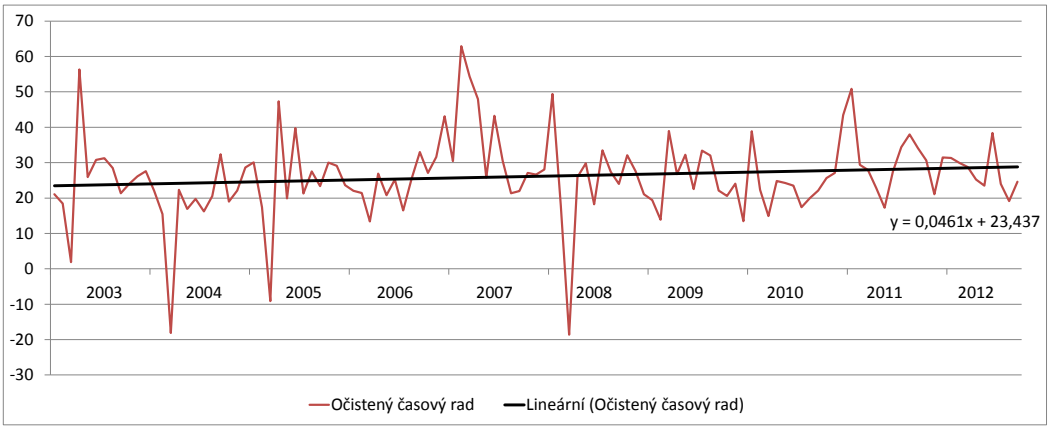


Obrázok 7.9: Porovnanie pôvodného a očisteného časového radu

7.2.5 Odhad trendovej zložky

Jednoduchá lineárna regresia

Časový rad bol očistený od sezónnej zložky. V nasledujúcom kroku budeme hľadať trendovú zložku, pričom sme sa rozhodli pre lineárny trend. Analýzou reziduí boli v očistenom časovom rade identifikované 3 vplyvné body, ktoré vzhľadom k svojej atypickej hodnote (extrémne počty požiarov v mesiacoch marec 2006, apríl 2009 a marec 2012) boli z regresnej analýzy (viď. kapitola 4.4) odstránené.



Obrázok 7.10: Očistený časový rad s regresnou priamkou (odhad trendovej zložky)

Na obrázku 7.10 je zobrazený očistený časový rad s preloženou regresnou priamkou. Rovnica priamky bola určená pomocou programu Microsoft Excel a má tvar $\hat{y}_t = 23,4 +$

$0,05t$, kde $t = 1, 2, \ldots$. Z toho môžeme odhadnúť, že počet požiarov má jemne stúpajúci charakter, ale túto domnienku musíme najprv potvrdiť.

Pomocou t-testov bude overená stabilita modelu. Testujeme hypotézu $H_0 : \beta_i = 0$ oproti $H_A : \beta_i \neq 0$, kde $i = 1, 2$.

Koeficient	**Test**	**p-hodnota**
β_0	t-test	$\ll 0,001$
β_1	t-test	0,118

Tabuľka 7.5: Overenie stability regresného modelu

P-hodnota pre koeficient β_1 je 0,118, takže na hladine významnosti 0,05 nulovú hypotézu nezamietame, parameter β_1 nie je štatisticky významný a môžeme ho z modelu vypustiť. Koeficient β_0 je štatisticky významný ($p - hodnota \ll 0,001$) a nemôžeme ho z modelu vypustiť, rovnica regresnej priamky má potom tvar $y = 23,4$. Regresný model sme zmenili na $y = \beta_0$, preto by sme mali znova vykonať odhad metódou najmenších štvorcov pre nový model. Keďže b_1 môžeme považovať za nulové, potom sa hodnota b_0 bude rovnať priemeru. Preto trendovú zložku časového radu jednoducho odhadneme pomocou priemeru $\hat{y}_t = \bar{y} = 26,2$

Odhad trendovej zložky pomocou priemeru

Vzhľadom k tomu, že očistený časový rad považujeme za výber z normálneho rozdelenia, môžeme určiť 95% intervalový odhad jeho priemeru (pás spoľahlivosti) a 95% intervalový odhad jeho individuálnych hodnôt (pás predikcie). Jednotlivé intervalové odhady sa spočítajú ako

$$P(\bar{y} - \tfrac{s}{\sqrt{n}} z_{0,975} < \bar{Y} < \bar{y} + \tfrac{s}{\sqrt{n}} z_{0,975}) = 0,95,$$

$$P(\bar{y} - s \cdot z_{0,975} < Y < \bar{y} + s \cdot z_{0,975}) = 0,95,$$

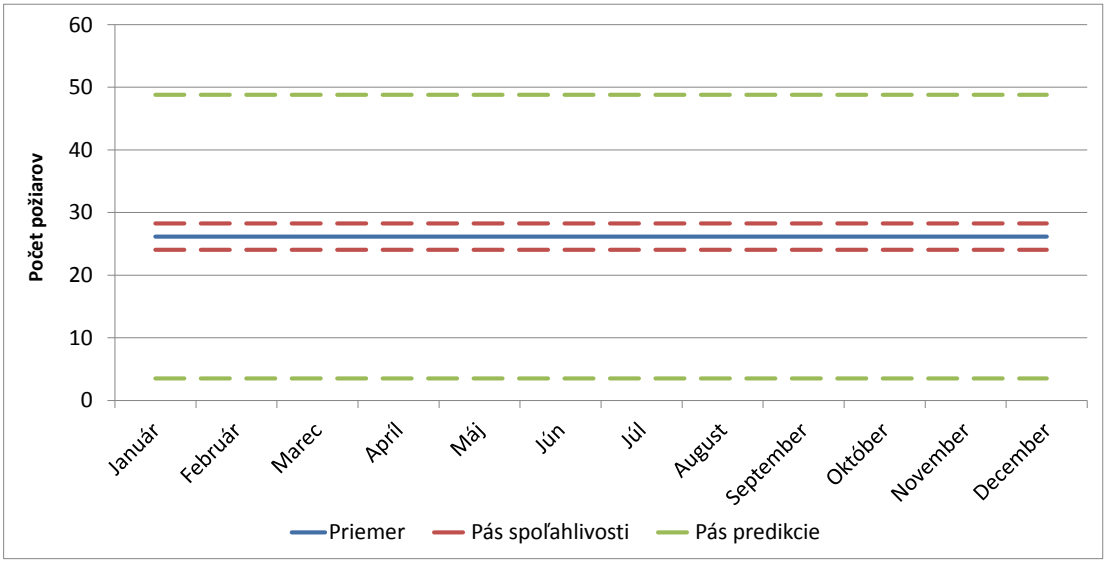
kde $\bar{y}$ je priemer očisteného časového radu, s je smerodatná odchýlka očisteného časového radu, n je rozsah výberu (počet všetkých analyzovaných mesiacov) a $z_{0,975}$ je 0,975 kvantil normovaného normálneho rozdelenia.

V našom prípade majú intervalové odhady tvar

$$P(24,1 < \bar{Y} < 28,3) = 0,95,$$

$$P(3,5 < Y < 48,8) = 0,95$$

a ich grafické zobrazenie je na obrázkoch 7.11 a 7.12.



Obrázok 7.11: Odhad trendovej zložky (s pásom spoľahlivosti a predikcie)

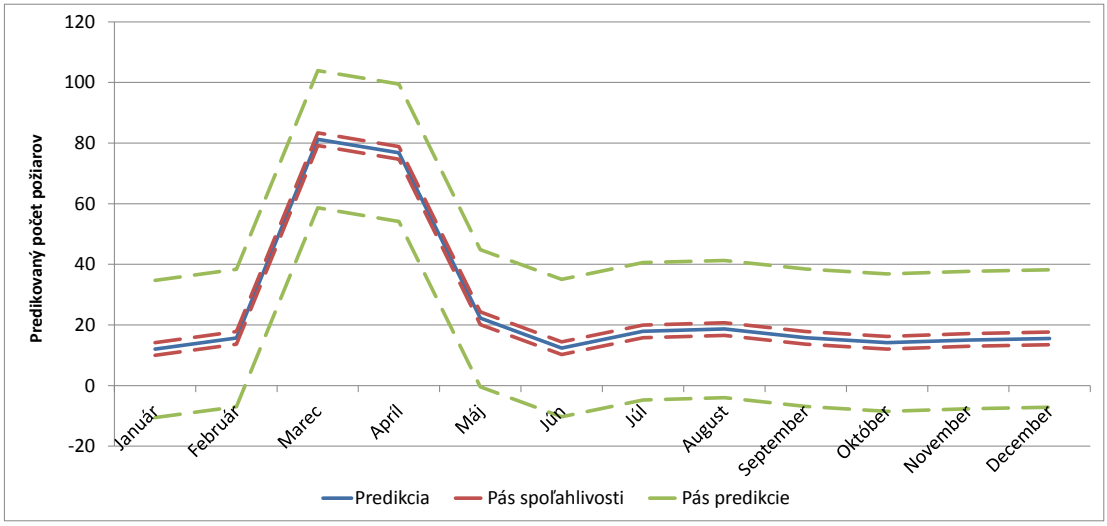
Pre lepšie odhadovanie počtu požiarov zahrnieme do modelu sezónnu zložku S_t.

$$\hat{Y}_t = T_t + S_t = 26,2 + S_t$$

Konkrétne hodnoty sezónnej zložky zobrazuje tabuľka 7.6.

Mesiac	**Sezónny faktor**
Január	-14,1
Február	-10,4
Marec	55,1
Apríl	50,6
Máj	-3,9
Jún	-13,8
Júl	-8,3
August	-7,5
September	-10,4
Október	-12,0
November	-11,1
December	-10,6

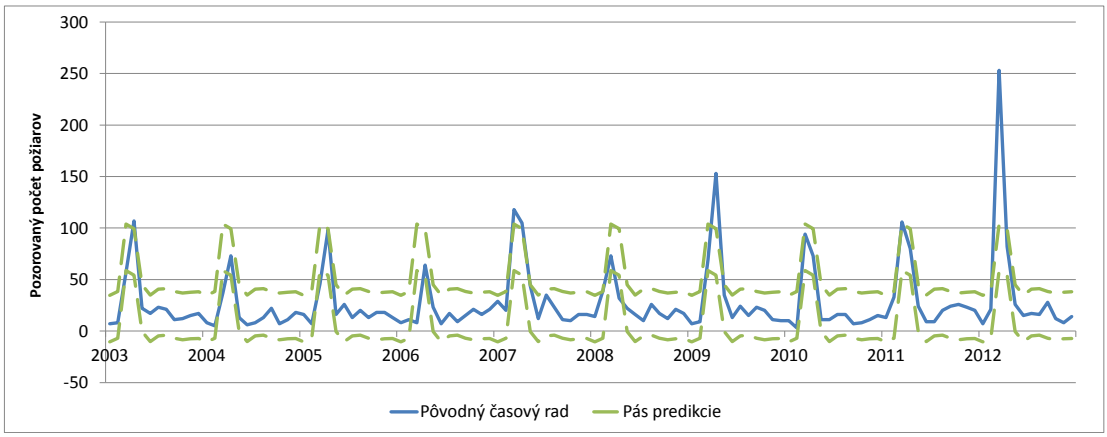
Tabuľka 7.6: Hodnoty sezónneho faktora pre jednotlivé mesiace



Obrázok 7.12: Predikcia počtu požiarov so zahrnutým sezónnym faktorom (s pásom spoľahlivosti a predikcie)

Na obrázku 7.12 vidíme, že pás spoľahlivosti aj predikcie kopíruje odhadovaný počet požiarov (súčet priemeru a sezónnej zložky). Napríklad v mesiaci marec môžeme do budúceho roka očakávať 60 - 100 požiarov.

Overenie kvality predikcie je možné pomocou analýzy doterajších údajov vzhľadom k pásu predikcie. Väčšina pozorovaných hodnôt by teda mala ležať v páse predikcie.



Obrázok 7.13: Overenie kvality predikcie požiarov

Tento odhad vývoja počtu požiarov môžeme považovať za spoľahlivý, pretože v analyzovanom období pás predikcie nepokryl iba 10 mesiacov zo 120, t.j. v 92% prípadov sa podarilo spoľahlivo predpovedať počet požiarov v danom mesiaci. Na záver tejto časti môžeme skonštatovať, že počet požiarov má počas 10-tich rokov ustálený charakter a nevykazuje žiadne nárasty resp. poklesy.

8 Záver

Cieľom bakalárskej práce bolo oboznámenie sa s metódami časopriestorovej analýzy v praxi.

Práca pozostávala z dvoch častí. V prvej polovici boli teoreticky opísané štatistické metódy, slúžiace hlavne k analýze kategoriálnych dát. Praktická časť sa venovala analýze požiarov, ktoré vznikli v katastrálnych územiach obcí patriacich do okresov Spišská Nová Ves a Gelnica medzi rokmi 2003 - 2012.

Zistilo sa, že počet požiarov počas týždňa je výrazne vyšší v sobotu. Napriek tomu, pomocou analýzy rozptylu sa mohlo vysloviť tvrdenie, že počet požiarov medzi jednotlivými dňami týždňa sa štatisticky významne nelíši. Ďalej pomocou kontingenčných tabuliek sa hľadala závislosť medzi príčinou požiaru a zásahovým obvodom. Spomínaná závislosť síce existuje, ale je pomerne slabá. Nakoniec sa skúmal vývoj počtu požiarov počas roka pomocou časových radov. Zistením, že počet požiarov má počas celého analyzovaného obdobia približne konštantný charakter bola praktická časť práce ukončená.

Na výpočty a grafické zobrazenia boli využívané funkcie štatistického softvéru „R“, Microsoft Excel 2007 a Statgraphics Plus 5.0.

V tejto baklárskej práci som získala reálnu predstavu ako skúmať veľký počet dát a vybrať správnu štatistickú metódu. Táto práca by sa dala rozšíriť napríklad o kartogramy zobrazujúce počty požiarov v jednotlivých obciach. Na analyzované dáta by sa dalo pozrieť z iného uhla, skúmať ďalšie závislosti (napr. vývoj požiarov počas dňa, závislosť škody spôsobenej požiarom na príčine požiaru) alebo využívať iné štatistické metódy (napr. pri analýze trendovej zložky časových radov využiť exponencionálne vyrovnávanie). Táto zaujímavá téma by mohla byť námetom pre diplomovú prácu.

9 Literatúra

[1] LITSCHMANNOVÁ, Martina. *Úvod do statistiky [online].* Ostrava, 2011. Dostupné z: http://mi21.vsb.cz

[2] BRIŠ, Radim a Martina LITSCHMANNOVÁ. *STATISTIKA I. pro kombinované a distanční studium [online].* Ostrava, 2004. Dostupné z: http://homel.vsb.cz/ bri10

[3] SHAPIRO, S.S. a M.B. WILK. *An Analysis of Variance Test for Normality (Complete Samples) [online].* 1965, 599 - 611.

[4] ANDĚL, Jiří. *Matematická statistika.* Praha, 1978.

[5] ŘEZANKOVÁ, Hana. *Analýza dat z dotazníkových šetření.* druhé vydání. Praha, 2010. ISBN 978-80-7431-019-5.

[6] PECÁKOVÁ, Iva. *Statistika v terénních průzkumech.* 2. dopl. vyd. Praha, 2011. ISBN 978-80-7431-039-3.

[7] HANČLOVÁ, Jana a Lubor TVRDÝ. *Úvod do analýzy časových řad [online].* Ostrava, 2003. Dostupné z: http://gis.vsb.cz/pan-old/Skoleni_Texty/TextySkoleni/AnalyzaCasRad.pdf

[8] CYHELSKÝ, Lubomír a Eduard SOUČEK. *Základy statistiky.* Praha, 2009. ISBN 978-80-7408-013-5.

[9] FLOREKOVÁ, Ľubica a Marta BENKOVÁ. *Štatistické metódy [online].* Košice, 2006. ISBN 80-8073-527-1. Dostupné z: http://www.scribd.com/doc/131302286/Štatisticke-metody

Prílohy

Príloha A

	Pondelok	Utorok	Streda	Štvrtok	Piatok	Sobota	Nedeľa
$\bar{x}$	2,8	3,0	3,5	3,9	3,6	4,1	3,1
min	1,7	1,7	1,5	2,4	1,9	2,6	1,5
$x_{0,25}$	2,1	2,1	2,9	2,7	3,0	3,2	1,9
$x_{0,5}$	2,6	2,8	3,5	3,8	3,5	4,0	3,2
$x_{0,75}$	3,5	4,2	4,1	4,7	4,0	4,6	3,9
max	4,5	4,6	5,5	6,0	5,1	6,8	5,1
IQR	1,4	2,1	1,2	1,9	1,0	1,3	1,9
s^2	1,1	1,4	1,2	1,5	0,9	1,4	1,7
s	1,0	1,2	1,1	1,2	0,9	1,2	1,3

Tabuľka A.1: Súhrnné štatistiky pre zásahový obvod Spišská Nová Ves

	Pondelok	Utorok	Streda	Štvrtok	Piatok	Sobota	Nedeľa
$\bar{x}$	2,4	3,1	3,0	3,3	3,8	4,3	3,4
min	0,0	0,9	0,0	0,4	1,7	2,2	1,3
$x_{0,25}$	1,2	2,2	1,4	2,2	2,6	2,6	2,2
$x_{0,5}$	2,2	3,0	2,2	3,9	3,7	3,2	2,6
$x_{0,75}$	3,4	4,3	4,6	4,3	4,7	5,6	4,1
max	4,7	6,0	6,9	4,7	6,9	8,2	8,6
IQR	2,3	2,2	3,2	2,2	2,2	3,0	1,9
s^2	2,3	2,8	5,2	2,1	2,7	5,1	5,0
s	1,5	1,7	2,3	1,5	1,6	2,3	2,2

Tabuľka A.2: Súhrnné štatistiky pre zásahový obvod Gelnica

	Pondelok	Utorok	Streda	Štvrtok	Piatok	Sobota	Nedeľa
$\bar{x}$	4,6	4,4	4,5	5,0	4,5	6,5	4,4
min	2,2	2,2	1,1	2,9	2,6	3,7	1,8
$x_{0,25}$	3,4	3,4	2,3	3,4	3,4	5,3	2,8
$x_{0,5}$	4,8	4,2	3,7	4,4	4,8	6,1	4,4
$x_{0,75}$	5,8	5,8	6,5	6,6	5,1	7,4	6,4
max	7,0	6,6	9,9	8,1	7,7	9,9	6,6
IQR	2,4	2,4	4,2	3,2	1,7	2,0	3,7
s^2	3,0	2,4	8,0	3,9	2,2	3,9	3,7
s	1,7	1,6	2,8	2,0	1,5	2,0	1,9

Tabuľka A.3: Súhrnné štatistiky pre zásahový obvod Krompachy

Príloha B

Číslo	Názov príčiny požiaru
1	Ďalšie sledované príčiny
2	Deti a choromyseľné osoby
3	Nedbalosť a neopatrnosť dospelých
4	Nezistená
5	Porucha, nevyhovujúci stav vykurovacích telies, dymovodov a komínov
6	Prevádzkovo-technické poruchy
7	Samovznietenie, výbuchy s následným požiarom
8	Úmysel

Tabuľka B.1: Číselník príčiny požiarov

Príloha CD

Priložené CD obsahuje pôvodné dáta o požiaroch v súbore $povodne\_data.xlsx$.